THE UNIVERSITY
of EDINBURGH

Inference of Evolutionary and Ecological Processes from Reticulate Evolution in RNA Viruses

Gytis Dudas

Doctor of Philosophy

School of Biological Sciences

Institute of Evolutionary Biology

University of Edinburgh

2015

Table of Contents

Abstract

RNA viruses have the fastest evolutionary rates amongst protein-coding organisms on the planet. Ease of sequencing, advanced techniques of analysis and global health and economic concerns have all contributed to the recognition of RNA viruses as a robust research platform. Phylogenetic methods have been at the forefront of analytical techniques used to understand the dynamics of RNA viruses – during natural circulation in populations and in individual hosts, within epidemics, across species barriers and over billions of years that viruses have been around.

Most of the work presented in this thesis employs phylogenetic incongruity arising from reassortment and recombination to gain insights into the genomes and populations of RNA viruses.

Chapter 2 explores the selection regimes Ebola virus has experienced following a year of circulation in humans in West Africa, as well as its recent history. Chapter 3 investigates the extent of recombination in MERS-CoV, a novel human pathogen with an obscure epidemiology, which is suggestive of frequent co-infection of some hosts. Chapter 4, on the other hand, documents a pattern of non-intuitive linkage between some segments of the human-endemic influenza B virus genome and explores its potential to speciate. Chapter 5 builds upon chapter 4 and attempts to describe small-scale reassortment between two segments of influenza B virus and the overall migration patterns of influenza B virus in Scotland. Chapter 6 exploits the independence of segments of influenza D virus, a recently described cattle pathogen, and coalescent theory to disentangle the origins of this virus.

This thesis exemplifies the success of modern sequencing methods, which, together with the use of sophisticated analytical techniques, have uncovered a wealth of information hidden away in molecular sequences of RNA viruses. The work presented herein demonstrates how reticulate evolution can be exploited as a reliable, and sometimes indispensable, marker to improve inference of evolutionary forces in RNA viruses.

Lay Summary

RNA viruses are a group of rapidly changing pathogens, some of which can infect humans. Similar to organisms which reproduce sexually, some RNA viruses are capable of unlinking parts of their genome, which to a large degree can then evolve independently. Similar to mutations, which can preserve signals of past population size changes and geographic movements, network-like evolution within genomes can be exploited to learn more about RNA viruses in question. Chapter 2 investigates the changes that have accumulated within the genomes of Ebola viruses circulating in West Africa during the 2013-2015 epidemic. Chapter 3 shows that some isolates of MERS coronavirus, an enigmatic virus causing sporadic infections in the Middle East, has exchanged small parts of its genome with other MERS coronaviruses. Chapter 4 shows that correlations exist between histories of particular genome segments of influenza B virus, a relatively prevalent human pathogen. Chapter 5 expands upon chapter 3 by focusing on short-term evolution of influenza B viruses as well as their human-assisted movement within Scotland. Chapter 6 shows how the independent histories of unlinked genome segments of influenza D virus, a recently described cattle pathogen, can be used to argue that it is a recently introduced, rather than an old pathogen. Overall, the work presented in the thesis shows how exchange of genetic information between RNA viruses can be employed to learn about multiple aspects of their biology and epidemiology.

Acknowledgements

The work described in this thesis would have been impossible were it not for the help, support and guidance of my supervisor, Andrew Rambaut. His keen insight, agile mind and cool-headed wisdom have led me through the field of molecular evolution, phylogenetics and programming and taught me much about the black art that is Bayesian statistics. More importantly Andrew has taught me to question everything, especially my own work, which remains by far the most valuable lesson of all.

I owe thanks to my PhD siblings and kindred spirits drawn to the field of virus evolution. In alphabetical order they are: Trevor Bedford, Liam Brierley, Luiz Max Fagundes de Carvalho, Matthew Hall, Jessica Hedge, Emma Hodcroft, Andrew Leigh-Brown, Samantha Lycett, Lu Lu, Jayna Raghwani, Melissa Ward and Paul Wikramaratna. They have improved me as a young scientist through countless discussions at lab meetings, conferences and coffee breaks. I would also like to thank my former officemates Reuben W. Nowell, Manon Ragonnet-Cronin and Darren J. Obbard. The first years of my PhD would not have been anywhere near as fun nor enlightening without you.

I am forever grateful to the people of Ashworth who made the department by far the friendliest place to work in. This includes PIs, postdocs, PhD students and staff, past and present, who took time off from their extremely busy schedules to sit, drink coffee or beer and chat about pretty much anything as equals. There is not enough room on a single page to express how much I appreciated every single PhD student who was nearly finishing up their own theses when I first came to Ashworth and who were giants in my eyes, everyone in my cohort who went through the same troubles and joys as I have and of course all the subsequent generations of PhD students I saw since, whose confusion and blunders betray the humble beginnings of those who will eventually become giants themselves.

I would also like to thank two overlapping groups of usual suspects by name. First, my fellow adventurers and persons of mayhem Richard Allen, Kevin Donnelly, Bryony Jackson and Maarit Mäenpää. Second, my loyal "bad" film night crew – Thomas Godfrey, Johann Hlina, Jisca Huisman, Ilkka Kronholm, Doris Reineke, Charlotte Repton, Elisa Schaum, James R. Westrip and Ashleigh Whiffin. And finally and definitely not last, María Fernanda Torres Jimenez, my fellow absinthe aficionado, a modern day conquistador of Europe, explorer of music and partner in crime.

As a side note the cover page of each of the seven chapters is embellished with a 3D protein structure of the major proteins encoded by members of the *Orthomyxoviridae* family of RNA viruses, of which influenza viruses are members. These are: PB1, PB2, P3 (all part of PDB accession 5D98), HEF (1FLC) of influenza C virus and NP (3ZDP), M1 (2Z16) and NS1 (3M53) of influenza A virus. Influenza viruses, especially the A and B kinds, have been a sort of *alma mater* to my foray into virus evolution research. The number of chapters amusingly corresponds to what I believe is the number of segments influenza viruses possessed ancestrally and the number that influenza C and D viruses still have to this day.

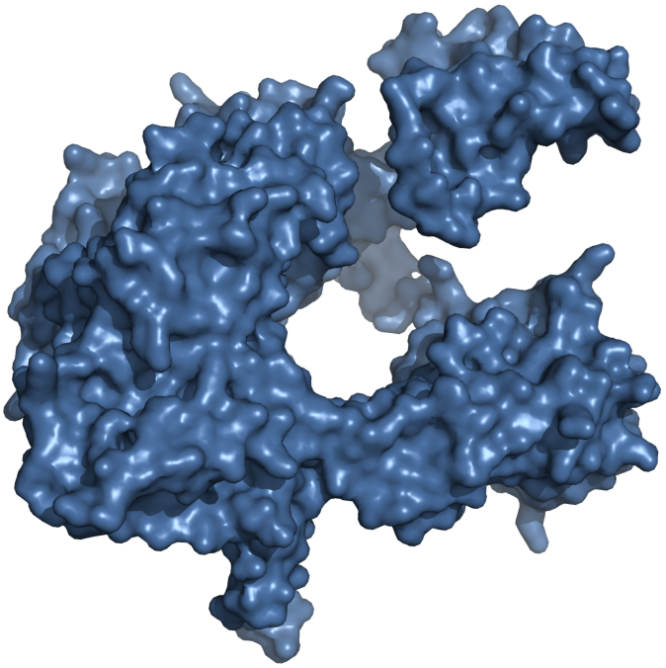
Declaration

I declare that this thesis was composed by myself and that all the work described within is my own. It has not been submitted for any other degree of professional qualification except as specified.

Gytis Dudas, 2015

CHAPTER 1

GENERAL INTRODUCTION



Then began I to thrive, | and wisdom to get,
I grew and well I was;
Each word led me on | to another word,
Each deed to another deed.

– Hávámal, *stanza 142*

Introduction to the biological system

Viruses

Viruses are the most numerically abundant group of organisms on the planet (Bergh et al., 1989). They are obligate intracellular parasites and employ a staggering variety of strategies to enter host cells, transcribe and translate the information encoded in their genomes and avoid the immune defenses of their host. The diversity of these strategies is accompanied by an array of genome organization plans. Current scientific consensus has settled on the Baltimore system (Baltimore, 1971) for classifying viruses, which is based on how messenger RNA is produced, whether it is produced from RNA or DNA and whether it is single-stranded (ss) or double-stranded (ds) and, if it is single-stranded, its polarity either positive (5'→3') or negative (3'→5'). The Baltimore system currently recognizes 7 groups of viruses: dsDNA, ssDNA, dsRNA, (+)ssRNA, (-)ssRNA, ssRNA-RT and dsDNA-RT, also known as groups I-VII. In addition to encoding their genetic information as different flavours of DNA or RNA, some virus groups also possess genomes split across physically unlinked RNA or DNA molecules (said to be segmented) or even virus particles.

Genetic diversity of viruses

Of all 7 groups of viruses in the Baltimore classification RNA viruses exhibit the highest evolutionary rates (Duffy et al., 2008). This is often attributed to the inherently high error rates of RNA polymerases, presumed to be caused by the lack of proof-reading activity (Steinhauer et al., 1992) that DNA polymerases possess. High error rates also put a limit on RNA virus genome sizes, due to deleterious mutation load. At the core of this phenomenon is the need to balance the generation of genetic diversity (for evading host defences or adapting to new hosts) and maintenance of functional proteins. Smaller genomes can tolerate deleterious mutations better, since fewer nucleotide sites mean there are fewer errors in each replicated genome.

In addition to mutation, some viruses undergo various forms of recombination. This ranges from occasional recombination of discrete parts of the genome to recombination that is extremely frequent and complicated. For example, influenza viruses (most

notoriously influenza A viruses) possess segmented genomes, where most proteins are encoded on single, physically unlinked RNA molecules (McGeoch et al., 1976) which themselves do not undergo recombination (Boni et al., 2008). This means that when two or more distinct lineages of these viruses co-infect a cell the progeny viruses can contain a combination of segments derived from multiple parental lineages. Because such exchanges of genetic material are discrete and invariably involve whole segments, it has been termed reassortment.

Coronaviruses, on the other hand, readily recombine during co-infection (Lai et al., 1985; Makino et al., 1986). This varies between relatively straightforward "switches" from one lineage template to another along a genome (Lai et al., 1985) to more complicated patterns, where fragments from distinct lineages alternate throughout the hybrid genome (Keck et al., 1988). Arguably the most complicated patterns of detectable recombination occur in Human immunodeficiency virus (HIV), a retrovirus. All retroviruses are pseudodiploid, where each virion contains two copies of the RNA genome (Beemon et al., 1974). The reverse transcriptase of the virus switches templates multiple times during reverse transcription of the genome from RNA to DNA, leading to observable recombination if genetically distinct genomes are co-packaged following co-infection (Jetzt et al., 2000).

Virus evolution

Much like other living organisms, virus evolution is driven largely by ecology. Many organisms experience and evolve in response to changes in their immediate environment, with availability of resources being one of the most important factors. Although arguments could be made that the availability of the raw building blocks for virus replication, such as ribonucleosides, deoxyribonucleosides, and amino acids, could affect how viruses evolve, one of the most powerful forces that shape virus diversity is the availability of hosts. Many RNA virus infections are ephemeral and require contact between an already infected host containing replicated progeny viruses and a näive or susceptible host (Villarreal et al., 2000). Few RNA viruses possess the ability to persistently infect their hosts, although well-researched systems exist such as HIV and hepatitis C virus. Even fewer viruses are known to be capable of altering the behaviour

of their host, examples of which would include baculoviruses that force their caterpillar host to seek out higher elevation prior to death and eventual shedding of virus on potential new hosts (Hoover et al., 2011) or the irritability of hosts induced by rabies viruses, which are transmitted via saliva. Given the rarity of the ability to persist within hosts or direct mechanisms to change host behaviour, RNA viruses often depend on existing contact networks between their hosts, either in the form of direct contact between individuals or indirect contact via vectors. The degree to which virus diversity is sculpted by their hosts also depends on the mobility of their host, which leads to population structure, and usually the kinds of immune responses the hosts can mount against the viruses that infect them. The integration of sequence data with current epidemiological, ecologcal and molecular phylogenetic theories has led to the birth of phylodynamics (Grenfell et al., 2004), a field of study aiming to understand what forces shape the genetic diversity of pathogens.

RNA viruses that infect vertebrates face a highly sophisticated adaptive immune system that is able to recall the last encounter with a phenotypically similar virus and mount a much faster response upon re-infection. The utility of such a strategy in most cases is limited, since most vertebrates do not live for a long time and thus any given virus lineage is unlikely to encounter the same individual during the individual's lifespan (Jong et al., 2007). Thus, even in the presence of adaptive immune responses most virus populations probably lose diversity via genetic drift rather than lack of adaptation (Holmes, 2008). Human influenza A viruses are one of the few RNA viruses where the opposite is true. The diversity of influenza A viruses infecting humans is almost exclusively shaped by viruses that have evolved to overcome the build-up of nearly population-wide immunity to the dominant strain, *i.e.* strong positive selection (Bedford et al., 2012). This is entirely down to the fact that humans are a long-lived host, where every individual will almost certainly encounter human-endemic viruses more than once in their life.

Viral emergence in humans

RNA viruses are over-represented amongst pathogens that sporadically emerge in humans, almost invariably from zoonotic (animal) sources. Although a random RNA virus drawn from the total global diversity of RNA viruses is unlikely to replicate in humans

successfully, the sheer diversity of RNA viruses means that a non-negligible number of these viruses will be able to make the zoonotic jump into humans. The nature of zoonoses and their eventual fate, however, involves a complicated interplay between ecology, evolution and epidemiology (Woolhouse et al., 2005; Parrish et al., 2008).

Some viruses, most famously influenza A viruses, successfully emerge in humans periodically. Two influenza pandemics of the 20th century, in 1957 and 1968, have been caused by influenza A strains seasonally circulating in humans acquiring a few genomic segments (invariably one of these has been a surface protein) from avian influenza A viruses through reassortment (Scholtissek et al., 1978a; Kawaoka et al., 1989). This was a consistent pattern until 2009, when an influenza virus generated through reassortment across 4 distinct genomic lineages from birds, swine and humans jumped into humans from pigs in Mexico and caused the first influenza pandemic of the 21st century (Smith et al., 2009).

However, the successful emergence of influenza A viruses is not always guaranteed – some genomic lineages, such as avian H5N1 and more recently H7N9, have provided a trickle of cases with high mortality and morbidity over the years in humans without ever achieving sustained human-to-human transmission (Qin et al., 2015). This pattern of stuttered transmission appears to be the case for the recently described Middle East respiratory syndrome coronavirus (MERS-CoV), for which non-hospital cases seem to be epidemiologically unlinked (Drosten et al., 2014) and there is ambiguous evidence of considerable human infection in the general population (Müller et al., 2015).

Other viruses, such as Ebola virus are readily transmissible, but are hugely dependent on the behaviour of the host. For example, the 2013–2015 Ebola virus epidemic in West Africa (still ongoing at the time of writing) has been very different from prior Ebola virus outbreaks in Central Africa in terms of scale and complexity. This has been attributed to differences in environment (isolated rural areas versus urban centres), burial practices and experience with managing Ebola virus disease (of which West African countries had none prior to the outbreak). Although Ebola virus is transmitted effectively from person to person, ultimately the virus is too pathogenic to become an endemic disease of humans and other primates.

Viruses covered in the thesis

Ebola virus

Ebola virus is a member of *Filoviridae*, a family that has become notorious for its ability to cause severe haemorrhagic fever in primates. Ebola virus, formerly known as Zaire ebolavirus, is a viral species within the Ebolavirus genus and was first identified in 1976 in the Democratic Republic of Congo (then known as Zaire) (Johnson et al., 1977) and has been an endemic pathogen of Central Africa since, causing devastating outbreaks in non-human primates and limited human outbreaks with a few, to a few hundred, cases in the Democratic Republic of Congo, the Republic of Congo and Gabon (Pourrut et al., 2005). At the beginning of 2014 the World Health Organization (WHO) declared an outbreak of Ebola virus disease first in Guinea, followed by Sierra Leone and Liberia. This was the first time this particular species of the virus was isolated in West Africa, although a very distant relative of Ebola virus, the Taï Forest ebolavirus, has been isolated in neighbouring Côte d'Ivoire (Le Guenno et al., 1995).

As the sequences for the West African outbreak, now the worst Ebola virus disease outbreak in recorded history, became available (Baize et al., 2014; Gire et al., 2014) several key questions emerged: where and when did the virus come from, how many species jumps into humans the virus had made and whether the scale of the outbreak made it particularly likely to adapt to humans. Chapter 2 investigates the effects of long-term genetic surveillance during the West African outbreak, compares it to early findings based on limited data from the same outbreak and contrasts the dynamics observed during prolonged human-to-human transmission with those of the reservoir. More importantly, due to concerns that have been raised about the potential for the virus to adapt, it also explores the evidence for positive Darwinian selection in the EBOV genome. The results from long-term surveillance show little evidence of positive Darwinian selection and instead suggest that lineages successfully transmitting in humans have, on average, fewer non-synonymous changes.

Middle East respiratory syndrome coronavirus

Coronaviruses are a group of three genera (Alpha-, Beta-, Gammacoronavirus) within the subfamily *Coronavirinae* of positive sense single stranded RNA viruses (King et al., 2011). They possess exceptionally large genomes amongst RNA viruses, speculated to be due to the presence of an exonuclease domain in the viral RNA polymerase (Gorbalenya et al., 2006). Some members of this group are frequent endemic respiratory pathogens of humans and cause mild symptoms ("common cold"), while others, such as severe acute respiratory syndrome and Middle East respiratory syndrome coronaviruses (SARS-, and MERS-CoV, respectively) have been at the centre of serious outbreaks in 2002-2004 (Peiris et al., 2003; Ksiazek et al., 2003) and 2012-2015 (Zaki et al., 2012) in humans with considerable case fatality rates. The outbreak of MERS-CoV is still ongoing at the time of writing. Both viruses are thought to have infected humans via an intermediate host, and ultimately appear to originate from bats (Li et al., 2005; Corman et al., 2014a). For SARS coronavirus the intermediate hosts were suspected to be palm civets (Guan et al., 2003), whereas for MERS-CoV there are consistent reports that a large number of dromedary camels across the Arabian peninsula have been infected with MERS-CoV at some point in their lives and only exhibit mild symptoms during infection (Müller et al., 2014; Corman et al., 2014b; Chu et al., 2014; Reusken et al., 2013, 2014).

Chapter 3 describes a multitude of tests that were performed on MERS-CoV to test for recombination. Recombination, the unlinking of neighbouring loci via template switching of the viral polymerase (Kirkegaard and Baltimore, 1986; Baric et al., 1987), has been extensively described in coronaviruses and their closest relatives (Lai et al., 1985; Makino et al., 1986; Keck et al., 1988; Kottier et al., 1995; Herrewegh et al., 1998). If not accounted for, it leads to false inference of Darwinian selection (Anisimova et al., 2003; Shriner et al., 2003) and can mislead phylogenetic methods (Schierup and Hein, 2000; Posada and Crandall, 2002). Many tests for recombination have been developed in the past, but often with relatively slowly evolving organisms in mind, where alleles segregate but are unlikely to arise *de novo*. Chapter 3 serves multiple purposes – it describes the performance of some recombination tests on temporally sampled sequences, identifies a pattern of rate heterogeneity in the MERS-CoV genome which appears to be shared with other coronaviruses and shows the presence of many

homoplasious sites most consistent with recombination. The latter finding is of key importance, since at present, the epidemiology of MERS-CoV is still poorly understood and the existence of a complicated recombination pattern implies that some host or hosts are frequently co-infected.

Influenza B viruses

Influenza B viruses, first discovered in 1940 (Francis, 1940), are the closest known relatives of influenza A viruses. Both are members of the *Orthomyxoviridae* family of segmented negative sense single stranded RNA viruses. Whereas influenza A viruses are known to have a diverse range of hosts, influenza B viruses have only been observed infecting humans, and seals through anthroponosis (Osterhaus et al., 2000; Bodewes et al., 2013), although experimental evidence suggests the virus can replicate in swine as well (Ran et al., 2015). This restricted host range limits the pandemic potential of influenza B viruses, as influenza A pandemics often result from seasonal human influenza A viruses acquiring antigenically novel surface glycoproteins from the animal reservoir, often birds.

Influenza B virus genomes are comprised of 8 physically unlinked genomic segments: polymerase basic 1 (PB1), polymerase basic 2 (PB2), polymerase acidic (PA), haemagglutinin (HA), nucleoprotein (NP), neuraminidase (NA), matrix protein (MP) and non-structural (NS). The first 3 segments (PB1, PB2 and PA) code for components of the viral RNA-dependent RNA polymerase (RdRp) complex, which is shared by common descent by all negative sense single-stranded RNA viruses (Li et al., 2015), although the number of proteins that make up the RdRp varies between different viral families. HA and NA code for the two main antigenic glycoproteins that are displayed on the surface of the virus. HA is used by the virus to bind host receptors and to release the viral genome in host cells via pH-induced membrane fusion. NA counters HA activity by cleaving host receptors in order to allow budding viruses to be released from infected cells. Nucleoprotein, encoded in the NP segment, coats the viral RNA genome and is another protein shared with all other negative sense single-stranded RNA viruses, although their homology between different groups has been questioned (Ruigrok et al., 2011). MP codes for two proteins, the matrix protein M1 and an ion channel M2. NS

codes for NS1 and NS2, which are used to modulate the host immune responses and to regulate the traffic of viral RNAs between the cytoplasm and the cell nucleus.

Both influenza A and B viruses circulate in humans seasonally and rely heavily on the steady accumulation of amino acid substitutions in their surface proteins, known as antigenic drift (Burnet, 1955), in order to evade the collective immunity of their host populations. Unlike influenza A viruses, which have circulated in humans as two antigenically distinct subtypes since the artificial introduction of one subtype in 1977 (Scholtissek et al., 1978b; Kendal et al., 1978), influenza B viruses, for the most part, have circulated as a single genomic lineage until the 1980s, when two genomic lineages of the virus emerged (Rota et al., 1990). Little is known about why this split occured, but antigenic differences between the two lineages have allowed both to co-circulate, albeit with contrasting epidemiological patterns (Vijaykrishna et al., 2015; Bedford et al., 2015).

Due to the presence of two distinct lineages, designated Yamagata and Victoria named after the strain names of representative viruses, as well as large amounts of circulating diversity, it has been possible to ascertain that influenza B viruses reassort fairly frequently when circulating in humans. Although much work has been done in determining the evolutionary history of influenza B virus genomes since their split into two lineages, there have been few studies (Hiromoto et al., 2000; Chen and Holmes, 2008) formalizing the observed patterns and even fewer looking into inter-dependencies between different segments. Chapter 4 thus focuses on influenza B virus reassortants that have combined segments from both Victoria and Yamagata lineages. The chapter presents methods for quantifying reticulate evolution between clades and assessing temporal co-divergence between two phylogenetic trees.

Influenza D viruses

Influenza D virus was first described in a study by Hause et al. (2013) after sequencing samples from a pig exhibiting influenza-like illness in Oklahoma. Although initially described as a divergent subtype of influenza C it was later shown that the divergence of this virus is comparable to that between influenza A and B viruses (Hause et al., 2013), which are considered to be separate genera within *Orthomyxoviridae*, and thus

influenza D virus and its relatives should be placed within their own genus. Subsequent surveillance identified cattle as a potential reservoir for the virus, and influenza D virus has since been consistently isolated from bovines in the United States (Hause et al., 2014), China (Jiang et al., 2014) and France (Ducatez et al., 2015). A recent discovery of a novel influenza virus in livestock raises important questions – has the virus been infecting cattle for a long time and surveillance missed it or is it a recent infection of cows?

Chapter 6 explores these two possibilities within the coalescent framework. Genomic data of influenza D virus are consistent with a common origin for all genomic segments of influenza D virus within the last 15 years. Although the data are consistent with a recent, rather than an ancient infection of cows, the geographic origins of the virus remain entirely obscure due to a long period of unsampled virus diversity. The influenza D virus story highlights an emerging view of influenza viruses as a clade of "generalist" viruses with a mostly unknown distribution and diversity in the wild, save for the known diversity of avian influenza A viruses. As such the influenza virus group poses a potential threat to livestock and humans.

Introduction to phylogenetic methodology

Methods of phylogenetic reconstruction

Phylogenetics as a field was transformed by the use of molecular sequences and development of distance-based algorithms for reconstructing phylogenies (Fitch and Margoliash, 1967; Saitou and Nei, 1987). Modern high-throughput sequencing methods have led to an avalanche of sequence information, most of which is easily accessible via online databases, such as GenBank (Benson et al., 2013). Phylogeny reconstruction methods settled on continuous time Markov chain (CTMC) methods to correct for unobserved multiple substitutions at the same sites (Jukes and Cantor, 1969). These nucleotide substitution models have increased in complexity since their introduction and have settled into their current form by 1986 (Tavaré, 1986).

CTMC models of sequence evolution also saw extensive use in the next generation of molecular phylogenetic methods, following additional improvements to account for

differences in evolutionary rates between sites (Yang, 1994). These methods were exemplified by explicit models of evolution and exhaustive searches through "tree space", a high-dimensional space where tree topology and branch lengths are parameters to be explored for a set of sequences in search of an optimal tree. The first successful method, following efficiency improvements (Felsenstein, 1981), was maximum likelihood (Felsenstein, 1973). Maximum likelihood aims to identify a single tree in tree space that maximizes the probability of observing the sequence data at hand. Bayesian phylogenetics, in contrast, arrived much later (Rannala and Yang, 1996), relies on posterior distribution approximation via Markov chain Monte Carlo (MCMC) methods (Metropolis et al., 1953) and aims to sample trees in proportion to their posterior probability and find the most probable distribution of trees that could have given rise to the sequence data.

Uncertainty in tree inference under maximum likelihood and Bayesian methods are fundamentally different. Maximum likelihood methods have come to depend on bootstrapping (Felsenstein, 1985a) to determine the uncertainty in tree topology reconstruction. This resampling technique samples alignment columns with replacement to create a number of pseudoreplicate datasets from which maximum likelihood trees are inferred. The number of times a particular clade is observed amongst pseudoreplicate trees is shown on the nodes of the actual tree. Other methods for constructing uncertainty intervals in a maximum likelihood framework depend on finding intervals of parameter values where the likelihood function drops below some threshold, although this is not a routine procedure in modern maximum likelihood-based phylogenetics. In Bayesian phylogenetics uncertainty in all parameters, not just tree topology, are recovered as part of MCMC sampling by default.

Molecular clocks

The molecular clock hypothesis states that the accumulation of changes is roughly constant with time, based on the observation that numbers of amino acid or nucleotide differences between two organisms are proportional to the amout of time that has passed since two organisms last shared a common ancestor (Zuckerkandl and Pauling, 1962; Margoliash, 1963). Molecular clocks have found themselves at the core of methods used to investigate infectious disease outbreaks, as they allow a phylogeny to be placed on

an actual timescale, with branch lengths proportional to time, rather than expected number of substitutions. This is of tremendous importance, since information about when common ancestors existed can be crucial (de Oliveira et al., 2006). In addition, the use of appropriately calibrated molecular clocks allows the phylogeneticist to extract additional information from sequence data (*e.g.* Dudas and Rambaut (2014)).

Initially molecular clocks were developed as a single rate for the entire phylogeny. This is sufficient in the absence of rate heterogeneity amongst lineages, but there are many cases where this is an unrealistic expectation. Well-sampled epidemics have a tendency to contain mildly deleterious segregating polymorphisms (Sharp et al., 2001). As the epidemic progresses, however, purifying selection removes deleterious polymorphisms and they are not observed in samples later in the epidemic. If sampling is very infrequent these polymorphisms can be missed and only alleles surviving purifying selection are observed. This effect can be observed even in densely sampled epidemics on internal branches, since they are the successful lineages that have at least two descendent sampled lineages and therefore should have experienced some degree of purifying selection. External branches or tips, however, represent the end points of lineages whose fate can be uncertain if they were isolated recently or, if they were sampled early, as lineages that are probably extinct. This difference in amounts of deleterious mutations sampled at the tips and missing mutations on internal branches because of purifying selection results in different evolutionary rate estimates on external versus internal branches (Pybus et al., 2007). At other times the opposite is true – in phylogenies of the HA segment of human influenza A viruses the series of internal nodes tracing the ancestry of the most recent tip to the root (called the trunk) are under extreme positive selection. This leads to the trunk having a higher evolutionary rate over any other class of branches in the phylogeny (Buonagurio et al., 1986; Ratner et al., 1989; Fitch et al., 1991).

To circumvent this issue a class of methods called relaxed molecular clocks were developed (Sanderson, 1997; Thorne et al., 1998). One of the most widely used subsets of relaxed molecular clocks are defined by specifying a lognormal or an exponential distribution from which evolutionary rates are drawn for each branch independently, such that rates on progenitor and descendant branches are uncorrelated (Drummond et al., 2006). The parameters of this distribution, such as mean and standard deviation, are then samped over the course of MCMC. This allows fine-scale reconstruction of

phylogenies and assessment of the magnitude of rate heterogeneity amongst branches. When sequence data have been sampled densely, especially in epidemic settings, the effects of sampling segregating deleterious polymorphisms can be mitigated somewhat by such relaxed clock models. Because uncorrelated clocks assume no autocorrelation between parent and descendent branches, abrupt changes in numbers of mutations on particular branches resulting from whatever process (*e.g.* host enzyme mediated hypermutation, or segregating deleterious mutations) can be accounted for and prevented from interfering with evolutionary rate estimates for the rest of the phylogenetic tree.

Alternative forms of relaxed molecular clocks include correlated clock models and local clocks. Correlated clock models work well when evolutionary rates are estimated over long periods of time. This is because properties of large populations that can affect evolutionary rate estimates (*e.g.* generation time, mutation rate, niche) usually change slowly and continuously through time. Correlated clock models vary in how exactly correlations between parent and descendent branches are induced: from minimizing changes in rate between branches to rates that are "inherited" from parent branches (Ho, 2009). Local clocks, on the other hand, are most often used when there is *a priori* suspicion that rate heterogeneity amongst lineages exists. These clock models are restricted to individual clades within the tree. Clades with different evolutionary rates can be designated in advance if there is suspicion that some properties of the entire clade can lead to significantly different evolutionary rates compared to the rest of the phylogeny *e.g.* different generation time (Smith and Donoghue, 2008) or changes in ecology (Worobey et al., 2014). Alternatively, random clades can be assigned separate clock rates as part of MCMC sampling procedure to identify clades with significantly different evolutionary rates (Drummond and Suchard, 2010).

The coalescent

The coalescent is a mathematical framework that describes the ancestry of a random sample of sequences taken from a population with discrete generations (Kingman, 1982). By running the genealogical process backwards the coalescent simulates genetic drift and so it has been at the core of methods attempting to recover ancestral population sizes from sequence data. Initially this was done by relying on simple parametric demographic models, which are used as tree priors. Tree priors serve as expectations

for the distribution of node times expected from different demographic processes. For example, under exponential growth most node times will be clustered towards the past, where coalescence occurs rapidly in a small population with node times becoming rarer closer to the present as coalescence rate decreases in a large population. Further developments have led to so called non-parametric tree priors which rely on measuring the coalescence rate directly from a phylogeny (Pybus et al., 2000). This class of methods, collectively known as the skyline plot family (Ho and Shapiro, 2011), saw extensive improvements, such as Bayesian smoothing (Drummond et al., 2005) of "relative genetic diversity" (the product of generation time and effective population size), explicit smoothing (Minin et al., 2008) and extending inference to multiple loci (Heled and Drummond, 2008; Gill et al., 2013).

Inference of trait evolution

There are many cases in phylogenetics where the phylogenetic tree itself is not of interest, but is required to ensure independence of observations (Felsenstein, 1985b). Sequence data are not independent observations of the evolutionary process, but rather a combination of independent changes acquired after lineages split from each other and shared changes that originally arose in a common ancestor and were inherited by sampled tips of a phylogeny. Each branch on a phylogenetic tree is an independent testable hypothesis of common descent. Mapping traits of interest on a phylogeny allows one to determine how many times and on which branches traits of interest have changed, given the observed states of the trait at the tips of the phylogeny. Examples of discrete traits that have been inferred in this way include geographic location (Slatkin and Maddison, 1989; Lemey et al., 2009a), host species of parasites (Smith et al., 1995; Weinert et al., 2012; Hall et al., 2013) and phylogenetic incongruence (Ward et al., 2013; Lu et al., 2014). State-of-the-art discrete trait inference methods have employed continuous time Markov chain matrices, much like the ones used for nucleotide substitution models, with arbitrary numbers of parameters determined by the number of states a given discrete trait can adopt.

Subsequent improvements to these methods have allowed the researcher to test for statistical significance of inferred rates. This is done via Bayesian stochastic search variable selection (BSSVS), a procedure whereby individual rates between discrete states

are turned on or off during MCMC. A prior is put on the number of rates that should be turned on, which by default is the minimum number of rates required to connect all the states (number of trait states minus one). This prior expectation allows one to calculate the ratio of prior odds of a given rate being on to the posterior odds (expressed as fraction of time a rate was turned on during MCMC versus when it was off) of the rate being on, which corresponds to Bayes factor (BF), which indicates a level of support that the rate actually exists.

The most sophisticated methods available now are even more explicit. Rather than estimating the rates of transition between discrete trait states and inferring their significance, researchers can specify predictor matrices *a priori* which quantify similarity or dissimilarity of discrete states. Over the course of MCMC individual matrices are either switched on or off, in a BSSVS procedure described earlier, which allows for explicit testing of significance. Additionally, individual matrices are scaled over the course of MCMC, such that the effect size of the predictor matrix can be assessed for effect size as well. This is a Bayesian form of generalized linear model (GLM) and has been used to great effect to uncover the underlying drivers of human influenza A virus subtype H3N2 global spread and inter-species transmission of rabies virus in bats.

Introduction to reticulate evolution

Reassortment and recombination

Most phylogenetic methodology relies on clonal evolution, *i.e.* that all loci being used contain information about a single tree topology whose parameters are to be inferred. Unequal evolutionary rates at different loci can be variously accounted for by accommodating differences in nucleotide frequencies across the alignment, mutational biases inherent in polymerases and evolutionary constraints. Problems arise, however, if the loci in question are unlinked by reticulation processes, such as recombination or reassortment. These processes transfer loci previously evolving in the same genetic background to a different background, resulting in independent evolutionary trajectories. This violates the assumptions of clonal evolution in phylogenetics, as well as any methods that rely on phylogenetic trees.

The two main processes that result in reticulate evolution are reassortment and recombination. Recombination is the exchange of genetic material at virtually any point along a genome. Unlike obligate sexual organisms, which usually possess dedicated cell machinery to facilitate recombination, recombination in RNA viruses is incidental. Usually the viral polymerase will disassociate itself from the template it is replicating and bind to another, resulting in the synthesis of a complementary sequence that is a hybrid of the two templates. This mode of recombination is known as template switching. Depending on whether the multiple templates shared a common ancestor at some point in the past (*i.e.* are homologous) or have evolved entirely independently up to that point, recombination can be further subdivided into homologous or non-homologous categories, respectively. Most recombination is likely to be homologous, simply because the nucleotides of the two recombining sequencing are expected to be a better match, on average, than between completely unrelated sequences. The increased nucleotide complementarity between homologous sequences thus makes it more likely that unrelated sequences will be brought into close contact with each other where template switching, the main mode of recombination in RNA viruses, can produce hybrid sequences. One obvious requirement for homologous recombination in RNA viruses is co-infection of the same cell with two or more related viruses.

Reassortment is unique to segmented viruses and results from the exchange of genomic segments between two or more related viruses co-infecting the same cell. It can be thought of as discrete homologous recombination, since many segmented RNA viruses appear to possess packaging signals to ensure the packaging of the correct number and kinds of segments in each progeny virion. Perhaps one of the most studied reassorting systems are influenza A viruses, due to the way viruses with potential to cause influenza pandemics are generated. Influenza pandemics occur when antigenically novel (to the collective immunity of humans) influenza A viruses capable of efficient person-to-person transmission arise. Due to the immense diversity of influenza A virus segments that code for proteins easily identifiable and targeted by the human immune system (haemagglutinin and neuraminidase) in birds viruses with pandemic potential are often produced via reassortment of human seasonal strains, which are already capable of efficient transmission between humans, and avian strains which possess surface proteins not previously seen by the human immune system. Although influenza viruses

are thought to undergo reticulate evolution exclusively via reassortment, recombination and reassortment are not mutually exclusively modes of evolution. Rotaviruses, which possess segmented positive sense double stranded RNA genomes are thought to undergo both reassortment between segments and recombination within segments.

Reticulate evolution detection

In some cases reassortment, such as influenza viruses, reticulate evolution is easy to detect by simply comparing the phylogenies of the segments. When there is sufficient information to reconstruct the phylogeny of each clonally evolving segment with sufficient confidence, any incongruence in tree topologies will arise from reassortment. Recombination detection, however, is usually more complicated. This is because unlike segments, which are clearly unlinked at their ends, the loci in recombining genomes are unlinked across arbitrary positions. As a result numerous methods have been proposed to detect reticulate evolution, each aiming to quantify the statistical properties of molecular sequences most affected by recombination. Recombination detection, at a minimum, requires that two or more recombining sequences are distinct. Additional factors, such as highly variable sites, the patterns, timing and rate of recombination can make recombination detection easier or more difficult. When the recombination tracts are short, there is little diversity in the alignment, recombination is frequent and has been ongoing for a long time the chances of successfully detecting recombination are minimal.

The most direct method of detecting recombination is inference of tree topology along an alignment suspected to harbour recombinant sequences. If there is strong support for two or more distinct and incongruent phylogenetic trees in the alignment it can be concluded that some loci of some sequences support an alternative tree topology than what has been inferred from neighbouring parts of the genome or the entire genome altogether. For example GARD, a tree-based recombination detection method, infers neighbor joining trees from fragments of the alignment, estimates the likelihood of the trees derived from those fragments and compares the resulting likelihood to the likelihood of a neighbor joining tree derived from the entire alignment. Tree-based methods, however, only work under certain circumstances. For one, the alignment has to contain polymorphic loci at a high density to ensure that even when the alignment

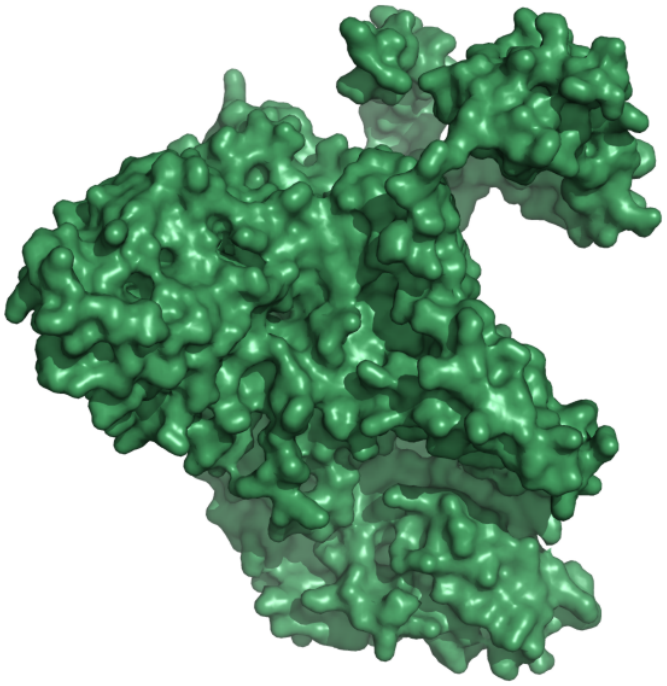
is separated into fragments there are sufficient numbers of informative loci to infer tree topology with any degree of confidence. Secondly, recombination will only be detected if multiple loci consistently point towards an alternative tree topology. This can only happen if recombination transfers distinct and long tracts of another genome, ideally on rare occassions such that both donor and recipient genomes usually evolve clonally. Depending on how the method is parameterised, additional complications can arise. In the presence of strong rate heterogeneity between sites, especially if it is systematic, methods like GARD will yield false positives, since both branch lengths (resulting from higher evolutionary rates) and topology contribute to the likelihood of a tree and cannot be differentiated easily.

A related method for detecting recombination is looking for excessive homoplasies. Homoplasies are changes in evolutionary history that have occurred more than once in a given phylogeny. The two alternative hypotheses for homoplasies are that they either evolved independently in different lineages or that they have been transferred from one lineage to another by recombination. In order to correctly infer homoplasies as such one usually needs a phylogeny which accurately describes a set of sequences. This is reasonable if recombination is rare, such that most changes inferred in lineages are due to independently acquired mutations. Another consideration in homoplasy inference is that similar patterns of positive selection can select for the same mutations in different lineages. However, in most cases such adaptations are mediated by amino acid changes, *i.e.* non-synonymous, rather than synonymous mutations. Recombination, for the most part, is expected to result in high levels of homoplasy across both synonymous and non-synonymous sites. Usually recombination detection methods relying on excessive homoplasies will use a phylogeny on which mutations are reconstructed and find a cut-off beyond which homoplasies are considered excessive. Recombination detection methods that exploit signals similar to excessive homoplasy detection, such as 3SEQ, attempt to find evidence of a specific sequence having been derived from two other sequences in the alignment. The difficulty in detecting recombination by inferring excessive homoplasies comes from the requirement for a phylogenetic tree, which does not describe the evolution of recombinant sequences well, and determining what number of homoplasies is considered as excessive, or in the case of triplet methods such as 3SEQ not having sampled sequences that could be one of the parents of a recombinant sequence.

The third class of recombination detection methods quantify the distances between sites displaying incongruent genealogies. If recombination is random, then sites further apart are more likely to have recombined in their history, simply because recombination is much more likely to happen over a longer stretch of nucleotides than over shorter stretches. There are several ways of quantifying incongruent evolution. Most often linkage disequilibrium (LD) is used, since it estimates the non-random association of sites along a sequence. LD is usually high for sites in close proximity and decays as the distance between the loci in question increases. Other methods for detecting incongruent evolution between loci involve looking for G4 sites. These are biallelic loci which contain four haplotypes *e.g.* if locus 1 contains alleles a and b, locus 2 contains alleles A and B, and all possible combinations or haplotypes thereof are observed (ab, aB, Ab and AB) then loci 1 and 2 are considered as G4 sites. It can be easily shown that the only way to get four haplotypes from two biallelic loci is if there has been a genuine repeat mutation (homoplasy) or recombination. Methods that employ LD decay usually either quantify the distance at which LD has decayed to some level or calculate statistics related to LD decay and compare the value observed for the actual alignment against permuted alignments, where site numbers (and therefore distances) have been randomised.

CHAPTER 2

EBOLA VIRUS EVOLUTION IN SIERRA LEONE



He must early go forth | whose workers are few,
Himself his work to seek;
Much remains undone | for the morning-sleeper,
For the swift is wealth half won.

– Hávámal, *stanza 59*

Analyses presented in this chapter were carried out on a more temporally extensive sequence data set, the results of which were published under the title "Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone" in *Cell* 161: 1516–1526. I analyzed the data and wrote this chapter. EBOV isolates were collected by Augustine Goba, Robert F. Garry and extensive teams at the Kenema Government Hospital in Sierra Leone, US Centre for Disease Control and Prevention and Tulane University. Samples were prepared for sequencing by Augustine Goba and Shanon L. M. Whitmer. Daniel J. Park led the project and assembled sequences.

Introduction

In December of 2013 an Ebola virus (EBOV) epidemic of an unprecedented proportion began in Guinea. Previously restricted to remote parts of Central Africa (Democratic Republic of Congo, Republic of Congo and Gabon) this was the first reported outbreak of Ebola virus disease in West Africa. Due to a combination of factors the outbreak spiralled out of control and eventually spilled over the borders of Guinea, most notably into Sierra Leone and Liberia.

The sequencing of Ebola virus isolates from the West African outbreak started in Guinea in March 2014 (Baize et al., 2014), 4 months after the presumed beginning of the outbreak. The virus was shown to be a recent and direct descendant of Central African EBOV strains (Dudas and Rambaut, 2014) rather than a divergent endemic lineage. When the virus was sequenced in Sierra Leone it became apparent that EBOV had continued spreading from person-to-person without any new introductions of the virus into human populations from the animal reservoir (Gire et al., 2014). The prolonged duration of the outbreak and an apparently elevated evolutionary rate within the Sierra Leone outbreak compared to the long-term rate (Gire et al., 2014) was widely misinterpreted as evidence of rapid evolution of the virus to adapt to its new human host (Olabode et al., 2015; Hoenen et al., 2015a).

Using an expanded set of sequence data from Sierra Leone encompassing an extra 4 months of virus circulation compared to the data in Gire et al. (2014) we explore the genetic changes that EBOV has undergone since its introduction into humans nearly

a year prior. Specifically, we focus on a question that has been raised before but not addressed – what form or forms of selection is EBOV undergoing during its extended human-to-human transmission in West Africa? We employ two amino acid change-dependent methods, a counting approach (O'Brien et al., 2009; Lemey et al., 2012) and an explicit parametric codon model (Goldman and Yang, 1994) to explore the selective pressures operating within the outbreak and within EBOV genomes.

Overview of EBOV biology

EBOV genomes are nearly 19kb long and composed of negative sense single stranded RNA, typical of *Filoviridae*, that codes for 7 structural genes – NP, VP35, VP40, GP, VP30, VP24 and L – separated by short intergenic regions (reviewed in Ascenzi et al. 2008). All negative sense single stranded viruses have a gene for N(P) (Ruigrok et al., 2011) and L (Li et al., 2015) proteins. Nucleoprotein, NP, coats the genomic RNA and L is the viral RNA-dependent RNA polymerase. The other genes – VP40, VP35, VP24, VP30 and GP – appear to be shared with other *Filoviridae*, but not with other negative sense single stranded RNA viruses. All genes except for GP code for single proteins. Transcriptional editing (Volchkov et al., 1995; Sanchez et al., 1996) produces 2 isoforms of GP – a highly expressed short secreted form and a longer membrane-bound form. Only the membrane-bound version of GP contains the mucin-like domain, a highly variable and glycosylated domain with little structure (Lee and Saphire, 2009). The soluble GP isoform has been suggested to be a decoy (Ito et al., 2001) for antibodies.

Little is known about the natural host of EBOV. It is strongly suspected, based on experimental infection (Swanepoel et al., 1996), serological and sequence evidence (Leroy et al., 2005), that the natural hosts of EBOV are bats. Some individuals from populations of three species of bats (*Hypsignathus monstrosus*, *Epomops franqueti*, *Myonycteris torquata*) have been shown to either have antibodies specific against Ebola virus or detectable RNA which is related to known Ebola virus isolates (Leroy et al., 2005). Between 1976 and 2015 there have been 12 recorded Ebola virus disease outbreaks caused by EBOV, 11 of them occurring in Central Africa: Democratic Republic of Congo, Republic of Congo and Gabon. Most Central African outbreaks have occurred in geographically isolated areas and were brought under control in humans relatively

rapidly, resulting in sequences from humans that are closely related. Because EBOV never had enough time to accumulate diversity before being extinguished, the collection of sequences from each human outbreak before 2013 are essentially the same identical or nearly identical sequence and represent a single draw of the viral diversity in the reservoir. The short sequences (ranging in size from about 300 nucleotides to 2100 nucleotides) generated by previous generations of sequencing methods, coupled with brief circulation periods of EBOV in humans and few sequences from the reservoir have led to the concept of "genetic stability" (Rodriguez et al., 1999), whereby EBOV is thought to evolve slowly, if at all. This terminology is misleading, as the long-term evolutionary rate in EBOV has been estimated to be around 7.06×10^{-4} (Carroll et al., 2013), 1.07×10^{-3} (Dudas and Rambaut, 2014) and 9.86×10^{-4} (Gire et al., 2014) substitutions per site per year, *i.e.* entirely consistent with previously described RNA virus evolutionary rates (Jenkins et al., 2002) and far from slow.

Methods

Sequence data and molecular phylogenetic analyses

Three datasets were constructed to represent three timescales of genetic surveillance of EBOV. 63 publicly available sequences representing the diversity of the virus sampled over long periods of time – starting with sequences from the 1976 outbreak, including sequences isolated from great ape carcasses, bats and human outbreaks with the exception of the DRC 2014 outbreak. The exclusion of sequences from the 2014 DRC outbreak was based on a known evolutionary rate anomaly associated with the outbreak (Lam et al., 2015), whereby the ancestral lineage leading up to the 2014 DRC outbreak experienced a markedly slower accumulation of substitutions than would be normally expected from EBOV. 14 sequences from West Africa were chosen to represent the current outbreak. 81 sequences analyzed by Gire et al. (2014) were reanalyzed here, representing the earliest epidemiologically-relevant and publicly available sequences. And finally 150 recently sequenced EBOV genomes from Kenema Government Hospital and 4 sequences from repatriated healthcare workers (UK, Italy and Switzerland) were combined with the 81 sequences from the early stages of the outbreak in Sierra Leone to give 235 sequences covering a large portion, both temporally and geographically, of the

epidemic in Sierra Leone. Synonymous and non-synonymous counts were mapped onto the molecular phylogenies using robust counting (O'Brien et al., 2009; Lemey et al., 2012) by specifying independent HKY nucleotide substitution models (Hasegawa et al., 1985) for all three codon position partitions. Substitutions in intergenic regions were modelled according to HKY with Γ_4-distributed rate heterogeneity (Hasegawa et al., 1985; Yang, 1994). A relaxed molecular clock with log-normally distributed rate categories (Drummond et al., 2006) and a non-parametric Bayesian skygrid (Gill et al., 2013) tree prior were used. An uninformative prior (Ferreira and Suchard, 2008) was used on the rate of the molecular clock. The demographic model for the long-term data was set to infer generation time-scaled effective population sizes at 20 time points with a cutoff 40 years ago, 20 time points and a cutoff 0.4 years ago for the reanalysis of Gire et al. (2014) data and at 50 time points and a cutoff at 1.2 years ago for 235 sequences.

We estimated dN/dS (ω) in EBOV genes NP, VP35, VP40, VP30, VP24 and L, using an implementation of the Goldman and Yang (1994) codon model in BEAST (Drummond et al., 2012). GP coding sequences were split into the mucin-like domain (GP_{MLD}), which encompasses amino acid residues 313–464 (Lee et al., 2008) starting from methionine, and the rest of GP ($GP_{\Delta MLD}$), due to concern that the GP_{MLD} is highly disorganized (Lee and Saphire, 2009) and thus under little constraint at the amino acid level. To date only linear epitopes in GP_{MLD} are known to be targeted by antibodies (Olal et al., 2012), due to its extensive O-, and N-linked glycosylation. We employed independent codon models for all 8 partitions, parameterised with independent strict molecular clocks due to limitations in model implementation. An uninformative prior (Ferreira and Suchard, 2008) was used on the evolutionary rate. Substitutions in the ninth partition, with concatenated non-coding intergenic regions, were modelled using HKY+Γ_4 (Hasegawa et al., 1985; Yang, 1994) model. The non-parametric Bayesian skygrid was used as the tree prior (Gill et al., 2013) for both long-term and current datasets, with the same inference parameters as above.

Results

Evolutionary rate within the West African outbreak is apparently higher

Substitution rate from the current dataset was estimated to be 1.280 (95% highest posterior density: 1.089 - 1.489) $\times 10^{-3}$ substitutions site^{-1} year^{-1} (figure 2.1). This is higher than the long-term estimate from a dataset spanning the entire sampled history of EBOV, at 1.064 (0.762 - 1.390) $\times 10^{-3}$ substitutions site^{-1} year^{-1}, but apparently slower than the estimate from the data of Gire et al. (2014), at 1.621 (0.931 - 2.414) $\times 10^{-3}$ substitutions site^{-1} year^{-1}. The difference in the amount of information on which to base rate inference is clearly reflected in the highest posterior densities (HPDs) of different data. The long-term rate estimate is based on 63 sequences sampled over nearly 40 years (figure 2.2). In contrast the dataset from Gire et al. (2014) covers 81 sequences sampled within 3 months of each other and the current data cover 235 sequences sampled over about 9 months (figure 2.3).

Split between West and Central African EBOV

The divergence between Central African (DRC) and West African strains has been variously dated to have occurred in 2002 (95% HPD 2000 - 2006) based on GP sequences alone (Dudas and Rambaut, 2014) and 2004 (95% HPD 2002 - 2006) based on whole-genomes (Gire et al., 2014). The latter result could have come about through imprecise estimates of the molecular clock rate due to the inclusion of circulating mildly deleterious variants in the data (Sharp et al., 2001). Figure 2.2 shows the phylogenetic tree of all EBOV strains sampled up to 2014 and a representative sample of West African strains. We estimate the split between Central and West African strains to have occurred around 2002 (95% HPD 1997 - 2006), which is mostly in line with previous estimates, albeit the uncertainty indicates a potentially older divergence date.

Circulation of deleterious variants

By separating inferred mutations in the tree into those that have occurred on external and internal branches *i.e.* those that lead to tips and those that do not, as

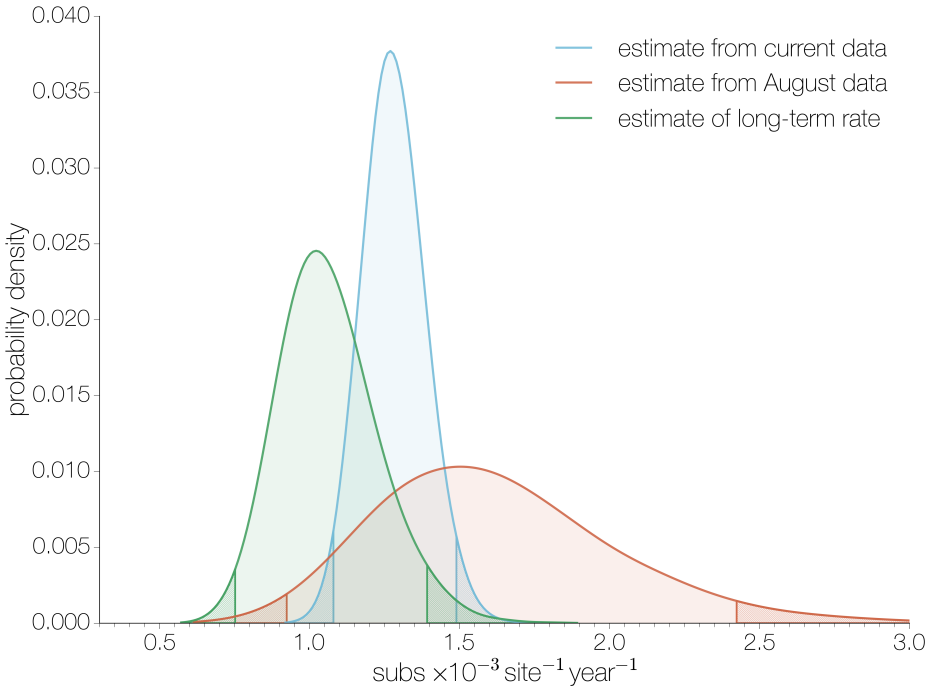


FIGURE 2.1: **Estimates of evolutionary rate from 3 datasets.** The 3 distributions show Bayesian posterior estimates of the evolutionary rate of EBOV from a dataset spanning nearly 40 years of evolution (green) in the reservoir, brief circulation period in humans (about 3 months) from Gire et al. (2014) (red) and the current data spanning nearly 9 months (blue). Although the new data have a lower estimated rate than the rate estimated from the data of Gire et al. (2014), the rate within the West African outbreak is still faster than the estimated long-term rate.

well as whether they result in synonymous or non-synonymous changes, we find that non-synonymous mutations have accumulated at a significantly faster rate on external branches (figure 2.4). This is not the case for synonymous mutations, which have accumulated at the same rate on internal and external branches.

No evidence of positive selection in West African EBOV

We find no strong evidence of positive selection in any of the EBOV coding sequences (figure 2.5). Over long periods we see strong evidence of purifying selection in all genes. Only the mucin-like domain of GP shows some evidence of positive selection over both long and short circulation periods (posterior probability $P(\omega > 1) = 0.9059$ and 0.9295, respectively), although estimates are still consistent with neutral evolution (ω 95% HPDs 0.880 - 2.385 and 0.673 - 17.409, respectively). All other genes have

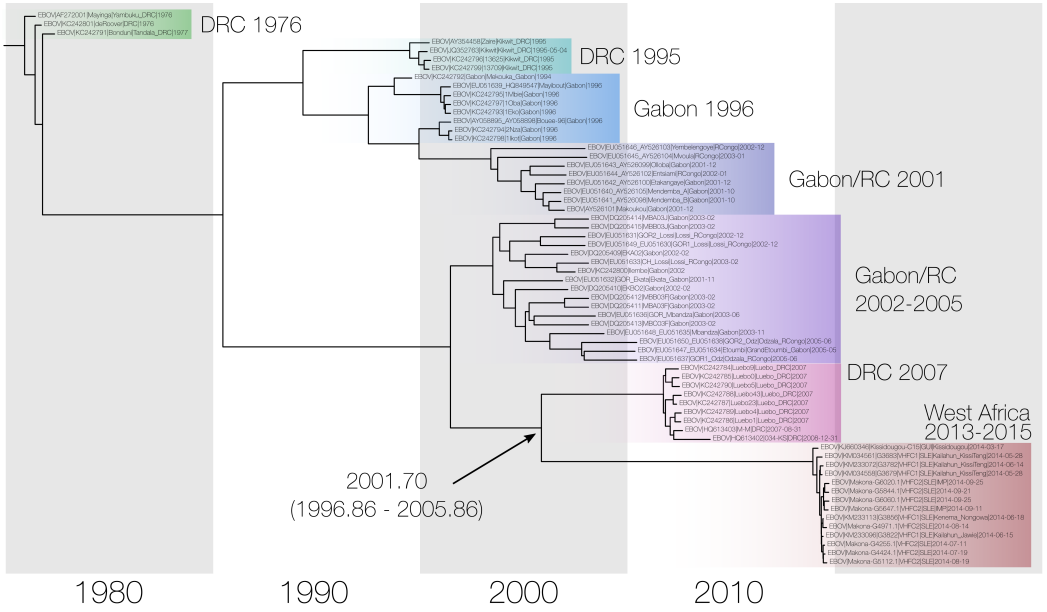


FIGURE 2.2: **EBOV evolution between outbreaks.** Maximum clade credibility (MCC) tree shows the relationships between sampled EBOV strains. The split and its estimated date between West and Central African EBOV strains is indicated with an arrow.

elevated dN/dS ratios (and uncertainty) within the West African outbreak compared to the long-term evolution of the virus, albeit NP, VP40 and L genes are still under apparent purifying selection within the West African outbreak.

Discussion

EBOV evolution at different timescales

Continued sequencing efforts have revealed a plethora of genetic diversity generated over the nearly two years of EBOV circulation in humans in West Africa. In contrast all information regarding long-term diversity of EBOV is derived from human outbreaks, which have all been sparked by a single genotype drawn from the reservoir. The genealogy of lineages "sampled" in this way is likely to trace EBOV lineages that have been under purifying selection in the reservoir during inter-outbreak periods. In contrast, lineages from within the West African outbreak are likely to be densely sampled and cover a significant portion of known cases, thus sampling extremely rare lineages that do not otherwise leave descendants. EBOV genomes with stretches of hypermutation have been isolated from the West African outbreak (Tong et al., 2015), with patterns

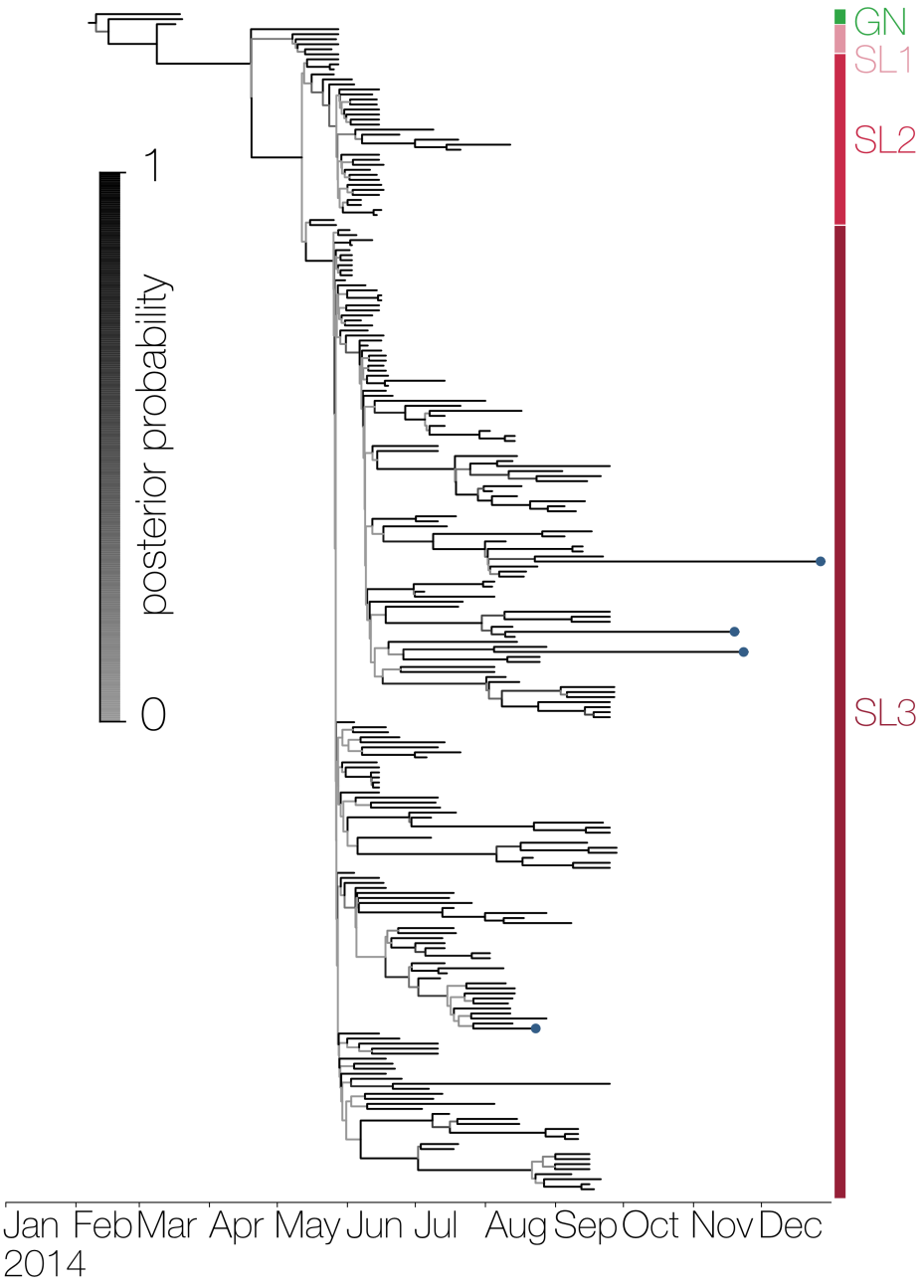


FIGURE 2.3: **Phylogenetic tree of West African strains.** Maximum clade credibility (MCC) tree shows the relationships between EBOV strains sampled during the West African outbreak. Bars on the right indicate clades using the nomenclature from Gire et al. (2014). Branches are coloured proportionally to their posterior probability. Repatriated cases are indicated by blue circles.

of hypermutation highly similar to those of double-stranded RNA-specific adenosine deaminases (ADARs) (Carpenter et al., 2009). Further complications are likely to arise due to differences in virus population sizes. Although virtually nothing is known about natural EBOV infection in the reservoir it is doubtful that significant similarities to human EBOV infection exist, especially since human infection is highly visible and

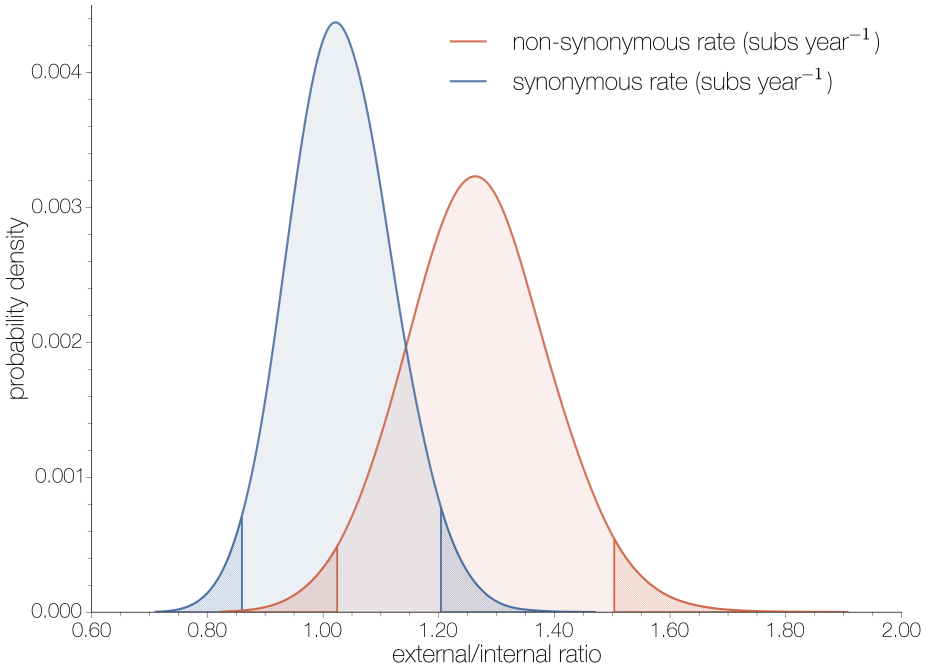


FIGURE 2.4: **Estimates of evolutionary rate from 3 datasets.** The 2 distributions show Bayesian posterior estimates of the ratio of rates between external and internal branches. Non-synonymous rates (red) differ significantly between external and internal branches, whereas synonymous rates (blue) do not.

eventually contained. The differences between virus population sizes in these two hosts, if they exist, are expected by theory to alter the efficiency of purifying selection, leading to faster fixation of viral genotypes in humans. In addition, differences in virus cell-to-cell generation times between hosts can dramatically affect the evolutionary rate, since viruses do not replicate outside of cells and thus cannot accumulate mutations.

Aside from uncertainties associated with virus replication in two presumably quite dissimilar hosts, it comes as little surprise that evolutionary rates estimated from viruses infecting those hosts give different results (figure 2.1). Evolutionary rates of influenza A virus lineages have been suggested to depend on host species (Worobey et al., 2014). The direction of this difference (higher rate in humans, lower in the reservoir), is consistent with predictions from population genetics theory and is a known issue in early outbreak investigations (*cf.* Rambaut and Holmes (2009) and Smith et al. (2009)). Studies claiming to have recovered a lower evolutionary rate within the West African outbreak have either recovered a rate that was not significantly different (Tong et al.,

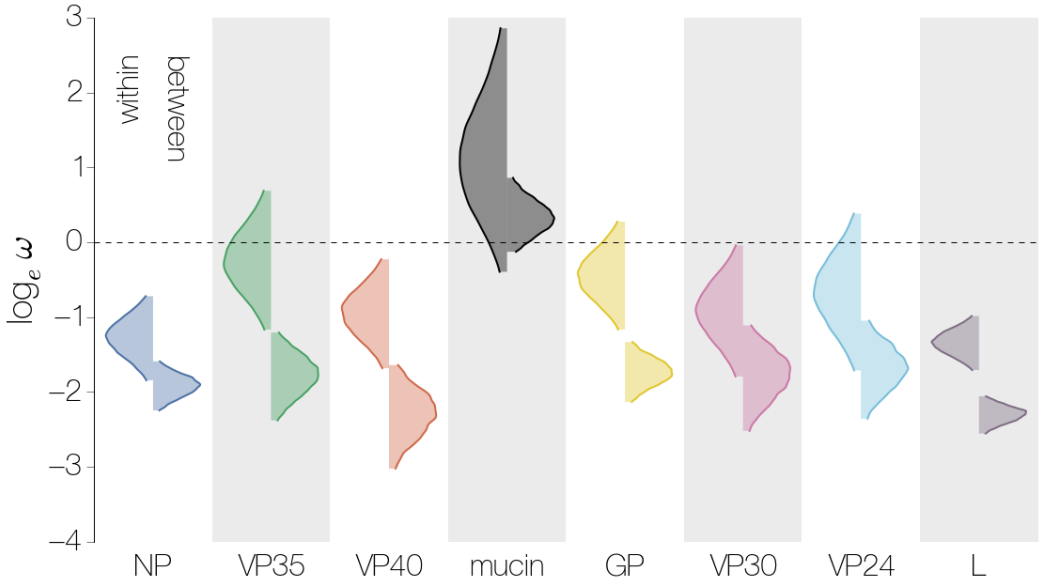


FIGURE 2.5: **Estimates of dN/dS in EBOV genes.** Violin plots of log(ω) estimates for coding regions of EBOV from the West African outbreak (left) and from all outbreaks (right).

2015) from the one described in Gire et al. (2014) or have estimated lower rates from sequences with randomized dates (Hoenen et al., 2015a). A re-analysis of the latter study (Hoenen et al., 2015b) found an evolutionary rate in line with previously published estimates. There is little significance to the magnitude of evolutionary rate in itself, although having estimates of it from an EBOV population sampled over a relatively large timescale is a considerable contribution to our knowledge of this virus and helps determine the temporal resolution at which future analyses would be operating.

Time scale-dependent evolutionary rate discrepancies are a known problem (reviewed in (Penny, 2005) and (Ho et al., 2011)). Almost invariably studies estimating evolutionary rates from sequences sampled over a short period of time or encompassing recently diverged lineages report higher rates than those estimated over long timescales. This problem is most acute in virus evolution research, where high rates in the present are irreconcilable with divergence estimates from external calibration which are thought to be more robust (reviewed in (Aiewsakun and Katzourakis, 2015)). The discrepancy has been variously explained by the action of purifying selection, which over longer timescales weeds out mildly deleterious variants segregating within populations, saturation, whereby constraint at the sequence level means that mutations are only tolerated and observed at specific sites and errors during sequencing or model misspecification

(Ho et al., 2005, 2011).

The improved rate estimates presented here allowed us to recover a more rigorous estimate for the divergence between Central and West African EBOV strains. Previous estimates of this date by Dudas and Rambaut (2014) relied solely on GP sequences whereas analyses by Gire et al. (2014) included all genomes (which evolve at an apparently elevated rate) available at the time. Given the effects of including genomes sampled densely over a short period of time during an active and prolonged epidemic (figure 2.1) we selected a small number of representative West African strains, which resulted in an older estimate for the split of West and Central African EBOV strains (2002, 95% HPD 1997–2006) than the one reported by Gire et al. (2014) (2004, 95% HPD 2002–2006) but also potentially older than that reported by Dudas and Rambaut (2014) (2002, 95% HPD 2000–2006). Although it is clear that the virus in West Africa arrived from Central Africa at some point within the last two decades, the route it took remains elusive. The lack of human cases along the path from Central to West Africa and lack of closely related viruses circulating in humans in Central Africa at the time discounts the possibility of direct human involvement in the migration. It is known, however, that the three species of bat from which Ebola virus-like sequences have been isolated – *Hypsignathus monstrosus*, *Epomops franqueti* and *Myonycteris torquata* (Leroy et al., 2005) – have a wide distribution across equatorial Africa and for the time being represent the most likely vehicle of viral migration.

Despite vast improvements to sequencing technology and breadth of sampling some limitations persist with regards to analyses presented here. Strains shown in figure 2.3 have been isolated across Sierra Leone and yet large portions of the West African EBOV tree are unresolved. This is in spite of full genome sequencing, which should improve the resolution of phylogenetic reconstruction by capturing more mutations, similar to how complete genome sequencing has transformed research into many bacterial pathogens (Biek et al., 2015). Additional complications arise because of sequencing methods – only consensus sequences are generated from deep sequencing of each patient sample. Simplifying this diversity down to single consensus sequences is misleading if virus population bottlenecks are sufficiently large at transmission and/or a considerable degree of genetic diversity accumulates within patients. Even if modern sequencing methods

were error-free, in the absence of the ability to recover the true linkage of individual mutations in sequencing reads which are usually too short for this, consensus sequences of diverse virus populations can contain artefactual homoplasies or can overlook genuine mutations. In addition, effects similar to incomplete lineage sorting, where phylogenetic trees are unable to capture the transmission tree (Pybus and Rambaut, 2009), are exacerbated when relevant genetic diversity within patients is ignored.

Selection regimes in human EBOV

Although the West African outbreak is unprecedented in scale, it is a fairly limited outbreak in terms of number of cases. As such, it is probably unlikely to ever encounter population-level immunity that influenza or other seasonally circulating and widespread RNA viruses encounter. The selection pressure to adapt to the human host, if it ever happened, probably occurred in the first case. Classic single amino acid polymorphisms determining host range have been described for influenza A viruses (Subbarao et al., 1993) and parvoviruses (Chang et al., 1992). At present it is impossible to determine whether such sites exist in EBOV due to lack of complete genome sequences from the reservoir. From what we know of other viruses, however, we expect EBOV to be under some degree of purifying selection.

Indeed, the pattern in figure 2.4 is most consistent with purifying selection. We observe that the most successful EBOV lineages represented by the internal branches of the phylogeny accumulate significantly fewer mutations per unit time than external branches, whereas synonymous mutation accumulation rates between internal and external branches are indistinguishable. Differences in non-synonymous rates indicate that successful variants which have left more than two sampled descendants have, on average, 1.53 fewer non-synonymous mutations per unit time than genotypes observed only once. This is strong evidence of deleterious mutation load and action of purifying selection which has been described previously for a wide variety of RNA viruses (Sharp et al., 2001; Pybus et al., 2007).

In addition to host adaptation and purifying selection there are additional selection regimes that are expected to operate during human infection. According to the WHO, EBOV has a variable incubation period in humans, which can range from 2 to 21

days, although it could be longer. There is ample time towards the upper tail of the incubation period distribution for the host to mount an adaptive immune response, with some complications arising through the biology of the virus. EBOV uses the GP protein to enter host cells. There are two main isoforms of GP: the highly expressed excreted and soluble form (sGP) and the rarer membrane-bound version found on the surface of virions. One of the major differences between the two isoforms other than their expression levels is the presence of the mucin-like domain, a disorganized sequence of amino acids with extensive O-, and N-linked glycosylation (Lee and Saphire, 2009), which is found only in the membrane-bound version. This comprises a highly efficient strategy of immune evasion, whereby host immune response is directed against the prevalent soluble GP isoform, which is otherwise of no use to the virus (non-structural), and the rarer but functional GP is heavily shielded by the mucin-like domain. We found no evidence of changes to the pattern of N-linked glycosylation in West African EBOV isolates, but did not test for changes in O-linked glycosylation due to difficulty in predicting this type of glycosylation from sequence data.

Further considerations of within-patient selection include the functions of accessory EBOV genes. Beyond the genes that are absolutely necessary to virus replication, packaging and transmission – the polymerase (L), nucleoprotein (NP), matrix (VP40) and surface (GP) proteins – EBOV has 3 additional genes with various functions: VP35, VP30 and VP24. VP35 is thought to be a type I interferon antagonist (Basler et al., 2000), much like the NS1 proteins of influenza viruses. VP30 appears to activate and regulate viral RNA transcription and VP24 is a protein of unknown function incorporated into the virus matrix. There are many examples of trait loss under altered ecological conditions (Lahti et al., 2009), thought to be mediated via relaxed selection pressure. It could be that some accessory EBOV proteins have evolved to optimize virus transmission in the natural host, but have lost, reduced or altered functions in humans.

Our estimates of dN/dS (ω) for coding regions of EBOV over long and short timescales (figure 2.5) indicate that the mucin-like domain shows strong evidence of positive selection, although the 95% HPDs cover 1.0, indicating that entirely neutral evolution cannot be discounted either. This result is consistent across long and short timescales.

Given the disorganized nature of the mucin-like domain and previous isolation of antibodies targeting only linear epitopes in this domain (Lee and Saphire, 2009), the universally high ω value probably reflects the ability of the immune responses of both the natural host and humans to target this domain, but also the ability of the virus to evade these responses via amino acid replacements. All other coding regions of EBOV within the West African outbreak exhibit ω values higher than those estimated for longer timescales in the reservoir. This might reflect a combination of factors, such as the aforementioned loss or alteration of gene function in a new host, but also population genetics effects such as reduced population size in humans. It is interesting to note that accessory genes VP35, VP30 and VP24, but also GP without the mucin-like domain appear to accumulate non-synonymous mutations at a rate consistent with entirely neutral evolution ($\omega \approx 1$). As expected, proteins analogous and/or homologous to proteins of other negative sense single stranded RNA viruses and thus presumably vital to viral fitness, *e.g.* matrix protein (VP40), polymerase (L) and nucleoprotein (NP) show evidence of purifying selection, in spite of relaxed selection. Additional considerations are the time scales involved in inference of selection. Ebola virus, like all RNA viruses, accumulate mutations fast compared to other organisms and with sufficient sequencing capacity it is theoretically possible to catch every minor variant that will go extinct within a given period of time. This concept of time-dependence of dN/dS-based inference of selection has been shown to occur in closely related bacteria (Rocha et al., 2006).

Future prospects

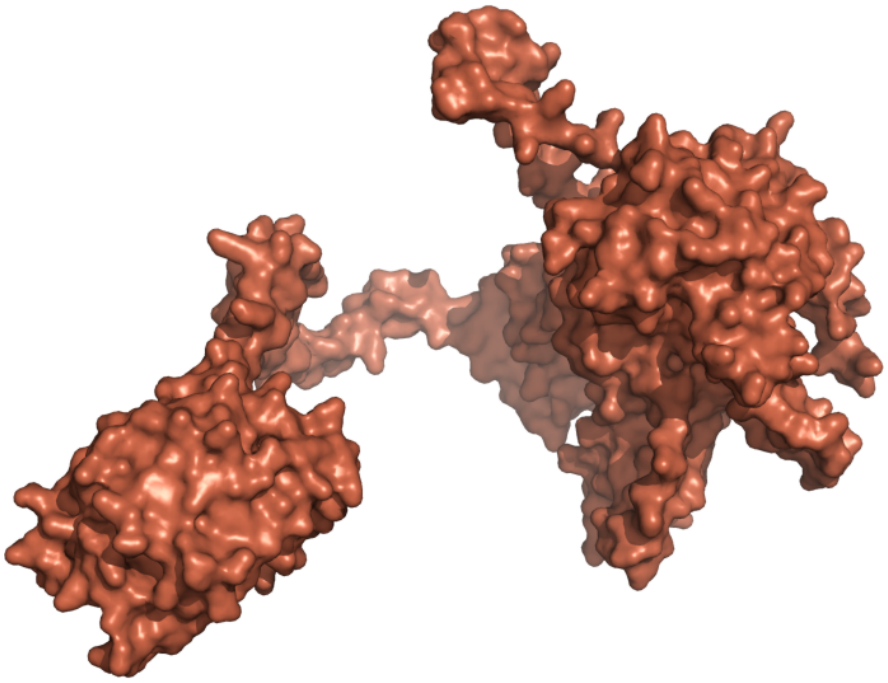
The continued emergence of EBOV in human populations in the future is a certainty. Historical trends would suggest that Central Africa will almost certainly continue to be at risk of EBOV outbreaks. In fact an outbreak of EBOV in the Democratic Republic of Congo did occur contemporaneously with the one in West Africa (Maganga et al., 2014) and was brought under control rapidly. The future prospect of another EBOV outbreak in West Africa at the moment is entirely unclear. It is still not known how the virus got to West Africa in the first place and it is unclear what epidemiological changes, if any, took place in the reservoir prior to EBOV zoonosis. Consistent surveillance and sequencing of the virus in the reservoir are key to understanding the epidemiology of

EBOV and the best, though remote, chance we will ever have of preventing future outbreaks.

Several key questions are likely to be addressed in the near future regarding the outbreak. First and foremost is the phylogeography of the virus. It has been clear from the beginning that country borders mean little to EBOV in West Africa, with indications of reciprocal virus migrations between the 3 countries. Second is intervention effectiveness. The initial international response to the crisis in West Africa failed, but it is uncertain how it failed exactly. One thing is certain, though, due to sequencing efforts, covering large portions of the epidemic in terms of numbers of case, geography and time, the West African EBOV outbreak is likely to become an epidemic studied for years to come.

CHAPTER 3

MERS-COV RECOMBINATION: IMPLICATIONS ABOUT THE RESERVOIR AND POTENTIAL FOR ADAPTATION



"A shadow on the wall," Varys murmured, "yet shadows can kill. And ofttimes a very small man can cast a very large shadow."

– George R. R. Martin, *A Clash of Kings*

This work has been published as **Dudas G, Rambaut A. 2015**. *Virus Evolution* 2(**1**):vev023 under the same title. I analyzed the data and wrote the manuscript.

Introduction

Recombination is an important process which expedites selection in many organisms (Muller, 1932) by unlinking loci. Different parts of recombining genomes have different evolutionary histories which, if not properly accounted for, can interfere with many genetic analyses, of which phylogenetic methods are amongst the most sensitive. Not accounting for recombination in phylogenetic analyses leads to incorrect (Schierup and Hein, 2000) and poorly supported genealogies (Posada and Crandall, 2002) and false inference of selection (Anisimova et al., 2003; Shriner et al., 2003).

With rising sequence availability during outbreaks of viral infectious disease, phylogenetic methods have been used to supplement our knowledge of epidemics in real time (Smith et al., 2009; Rambaut and Holmes, 2009; Lemey et al., 2009b; Drosten et al., 2013; Cotten et al., 2013, 2014; Drosten et al., 2014; Gire et al., 2014). For some outbreaks there is little reason to suspect recombination, *e.g.* negative sense single stranded RNA viruses are thought to recombine very rarely with recombination events observed over evolutionary time scales in whole clades, rather than continuously within populations during the circulation of these viruses (Chare et al., 2003). Observable recombination in RNA viruses requires that two conditions are met: that viruses from distinct lineages co-infect a host and that a mechanism for recombination exists. For example, even though influenza A virus co-infection is extremely common in birds based on genome segment reassortment patterns (Li et al., 2004; Dong et al., 2011; Lu et al., 2014), recombination is extremely rare or absent (Chare et al., 2003; Boni et al., 2010). This is thought to be because template switching (Kirkegaard and Baltimore, 1986; Baric et al., 1987), the main mechanism of recombination in RNA viruses, is mechanistically difficult for single stranded negative sense RNA viruses (see Chare et al. 2003), and for influenza A viruses has only been convincingly shown in cell culture under extreme conditions (Mitnaul et al., 2000). When the genomic architecture of a virus

is permissive to recombination, *i.e.* template switching occurs and is detectable, the extent of recombination is informative of co-infection and/or duration of infection.

Here we focus our attention on the Middle East respiratory syndrome coronavirus (MERS-CoV) (Zaki et al., 2012), a recent zoonotic infection with a relatively high case fatality ratio estimated to be around 20-65% (Cauchemez et al., 2014; Memish et al., 2013; Assiri et al., 2013). Most human infections with MERS-CoV are thought to be the result of contact with *Camelus dromedarius* L., the dromedary camel, which is the presumed host of the virus. MERS-CoV, much like Severe acute respiratory syndrome coronavirus (SARS-CoV), is likely ultimately derived from bats (Corman et al., 2014a). MERS-CoV, along with Murine hepatitis virus and SARS-CoV, belongs to the Betacoronavirus genus. Betacoronavirus, as well as two other genera (Alpha- and Gammacoronavirus) out of four within the subfamily *Coronavirinae* have been shown to recombine in cell culture, *in vivo* and in eggs (Lai et al., 1985; Makino et al., 1986; Keck et al., 1988; Kottier et al., 1995; Herrewegh et al., 1998). Additionally, a coronavirus lineage related to MERS-CoV which was isolated from bats appears to have recombined around nucleotide position 22000, near the gene encoding the spike (S) protein (Corman et al., 2014a). In this paper we show that although the genome of MERS-CoV contains considerable amounts of rate heterogeneity between genomic regions that can interfere with detection of recombination, we do nonetheless find evidence of sustained recombination that cannot be explained by rate heterogeneity alone. This has two important consequences: one is that care has to be taken when constructing phylogenetic trees of MERS-CoV as a single tree cannot accurately describe the complete history of all loci within a recombining genome. Secondly and more importantly, the observed rates of recombination in the MERS-CoV genome are evidence of a large number of MERS-CoV co-infections in some hosts which has implications for understanding the dynamics of the virus in the animal reservoir.

Methods

Data

Two main datasets were used for this study. One was composed of 109 MERS-CoV sequences from humans and camels. This dataset used in LDhat analyses is described later. Due to the large genome size of MERS-CoV most groups have sequenced amplicons covering large parts of the MERS-CoV genome, rather than the full genome. A subset of this dataset, composed of 85 sequences, was used for all other analyses. This dataset contained only those sequences which shared overlapping sequenced regions, since partitioning of the genome into discrete regions led to some strains not having any informative sequence for phylogenetic analyses, such as GARD. A list of all strains used, as well as GenBank accession numbers, where available, are at the end of appendix A.

Overview

Recombination leaves several characteristic clues in genomes:

- Alternative topologies (Robertson et al., 1995a,b; Holmes et al., 1999). In some scenarios, for example if there has been a single recombination event transferring a large number of divergent sites from one genetic background to another, it is possible to clearly identify recombining fragments based on phylogenetic incongruity. Recombination can be inferred by reconstructing two or more phylogenetic trees from a partitioned alignment and looking for topological incongruity between them. Strong support for at least two incompatible phylogenetic trees across well-defined breakpoints is usually the most convincing evidence of recombination.

- Excessive homoplasies (Maynard Smith and Smith, 1998). The transfer of genetic material from one genetic background to another will result in apparent repeat mutations in different parts of a phylogenetic tree. However, it is possible for the same locus to undergo mutation independently, especially if the locus in question

is under Darwinian selection. Detecting homoplasies alone is not sufficient to infer recombination, but should be demonstrated to occur in excess of expectation.

- Linkage disequilibrium (LD) decay (Miyashita and Langley, 1988). Linkage disequilibrium or LD is the non-random association of alleles at different loci. This is a statistic often reported for contemporaneous sequence data. In clonally (*i.e.* non-recombining) evolving organisms every allele is linked to every other allele in the genome and requires mutation to break linkage. In recombining organisms there is an expectation that LD will decay with distance between the loci, *i.e.* that loci further away from each other are more likely to be unlinked via recombination.

We test for each of these hallmarks of recombination in the MERS coronavirus genome using a combination of phylogenetic and linkage disequilibrium metrics. For a more detailed review of recombination detection methods see Posada et al. (2002).

Alternative topologies

We use the Genetic Algorithm for Recombination Detection (GARD) method (Kosakovsky Pond et al., 2006), as implemented in the software package HyPhy (Pond et al., 2005), to look for alternative tree topologies in sequence data. Briefly, the method compares a model where a single tree is derived from the whole alignment and alternative models where breakpoints are introduced into the alignment and phylogenetic trees are derived independently from the resulting fragments. The presence of recombination, especially if it is recent and concentrated in some parts of the alignment, will result in two or more phylogenetic trees fitting the data better than a single tree model. It is important to note that likelihood estimation also involves other parameters, such as branch lengths, not just topology. We use GARD under a GTR (Tavaré, 1986) substitution model with Γ_4-distributed rate heterogeneity amongst sites (Yang, 1994) on a dataset of 85 MERS-CoV sequences. GARD was run repeatedly until no more breakpoints could be identified in the resulting fragments.

In addition to this test, we run BEAST (Drummond et al., 2012) on partitioned coding sequences derived from the first well-supported breakpoint inferred by GARD. We

extracted the coding sequences from nucleotide positions 1-23722 and 23723-30126 (referred to as fragment 1 and 2, respectively) of MERS-CoV genomes. Independent HKY+Γ_4 (Hasegawa et al., 1985; Yang, 1994) nucleotide substitution models were specified for codon positions 1+2 and 3 and the analyses were run under an uncorrelated relaxed lognormal clock with an uninformative CTMC reference prior (Ferreira and Suchard, 2008) on the mean of the lognormal distribution for 100 million states, subsampling every 10000 states. The molecular clocks and trees of each genomic partition were either linked or unlinked, giving a total of 4 models. The CTMC reference prior is an uninformative prior which only requires the specification of the starting value and is recommended for cases where the evolutionary rate is not well established. We used the multi-locus skygrid (Gill et al., 2013) as the demographic model for all analyses. Path-sampling and stepping stone sampling (Baele et al., 2012) were used to calculate marginal likelihoods and test the fit of each of the 4 models, under default parameters. In addition, 4 similar analyses were set up, but with strict molecular clocks, in order to contrast the performance of relaxed molecular clocks.

Excessive homoplasies

Testing for recombination by looking for homoplasic mutations in phylogenetic trees requires that two conditions are met. One, that recombination is rare enough, so that there is sufficient phylogenetic signal to reconstruct the "correct" phylogeny otherwise known as the clonal frame (Milkman and Bridges, 1990). Two, that alternative explanations for homoplasic mutations can be dismissed with some certainty. There is no straightforward way of testing for the former, but the latter is usually dictated by the underlying biology. For example, repeat amino acid substitutions are a well documented response of influenza viruses and HIV to drug treatment (Gubareva et al., 2001; Tisdale et al., 1993; Boucher et al., 1993).

We employ two methods to test for excessive homoplasies. First, we use a maximum likelihood phylogeny inferred using PhyML (Guindon and Gascuel, 2003) under a GTR+Γ_4 (Tavaré, 1986; Yang, 1994) nucleotide substitution model to recover a single tree using a MERS-CoV dataset comprised of 85 sequences. We then reconstruct ancestral sequences at each internal node and identify the mutations that have taken place along each branch using ClonalFrameML (Didelot and Falush, 2007), which uses

a joint maximum likelihood algorithm to estimate the overall most likely sequence at each node (Pupko et al., 2000). This is in contrast to marginal maximum likelihood ancestral sequence reconstruction methods, which reconstruct the most likely ancestral sequence for each node individually. Mutations are then classified as either synapomorphies, shared variation derived via common descent or apparent homoplasies, shared variation derived from convergence, depending on how many times a given mutation has arisen in the phylogeny. The drawback of this method is that it necessarily conditions on a single tree with the highest likelihood.

We also employ BEAST (Drummond et al., 2012) to circumvent the limitation of conditioning the ancestral state reconstruction on a single tree. In addition to sampling various phylogenetic parameters from the posterior distribution BEAST is also able to map substitutions onto the branches of each MCMC-sampled phylogeny (O'Brien et al., 2009). This method is thus capable of estimating the posterior probability of a given mutation being synapomorphic or homoplasic by integrating over different tree topologies. Homoplasy analyses were performed on the concatenated coding sequences of MERS-CoV after partitioning the alignment into all three codon positions, each with an HKY nucleotide substitution model (Hasegawa et al., 1985) and no Γ-distributed rate heterogeneity amongst sites. Γ-distributed rate heterogeneity amongst sites was omitted from the model due to lack of implementation in BEAST. A relaxed uncorrelated molecular clock with lognormally distributed rates (Drummond et al., 2006) under a CTMC reference prior (Ferreira and Suchard, 2008) and the flexible multi-locus skygrid as the demographic model (Gill et al., 2013) were used. The MCMC chain was run for 100 million steps, sampling every 10000 steps.

Throughout the chapter we will refer to the number of branches that have experienced a given mutation as homoplasy degree. We define the homoplasy degree to be the number of times a given mutation has originated independently minus one. For example a homoplasy degree of 1 indicates that a mutation has occured on two different branches in the phylogeny. That is, we assume that one of the mutations has arisen through replication error, whereas the other has potential to have been introduced via recombination and thus can be thought of as excessive. Synapomorphies, on the other hand, are states that are shared by two or more taxa through common descent and

thus necessarily are those mutations that have occured exactly once in the phylogeny. They have a homoplasy degree of 0 in all figures.

Additional tests for recombination were also performed, namely the estimation of the pairwise homoplasy index (PHI) (Bruen et al., 2006) and the triplet test implemented in 3Seq (Boni et al., 2007).

Linkage disequilibrium decay

In the absence of recombination every allele should exhibit a high degree of linkage with other alleles in the genome. Under two extremes - clonal reproduction without recombination and free recombination - there is no correlation between LD and genomic distance and loci should be interchangeable. This is the basis of several non-parametric permutation tests for recombination that are implemented in the software package LDhat (McVean et al., 2002), which we used in combination with a dataset of 109 MERS-CoV genomes. Other, more complicated tests, such as composite likelihood methods, are also available but in our experience were incompatible with temporal sampling and rate heterogeneity.

Sequence simulations

To test the performance of some of the methods we simulated two sets of sequences. We use fastsimcoal2 (Excoffier et al., 2013) to simulate 10 replicate datasets that have the same dates of isolation and similar diversity to the MERS dataset with 85 sequences under no recombination.

Additionally, we use πBUSS (Bielejec et al., 2014) to simulate sequences down an MCMC-sampled phylogeny drawn at random from a linked-tree unlinked-clocks BEAST analysis described above. We modelled region-specific rate heterogeneity by simulating a 30kb "genome" and setting the molecular clock rate for the first 20kb to be 9.5×10^{-4} substitutions site^{-1}year^{-1} and the last 10kb to be 2.85×10^{-3}, 1.9×10^{-3} or 1.3×10^{-3} substitutions site^{-1}year^{-1}, corresponding to roughly 3-, 2- or 1.3-fold rate heterogeneity between the two parts of the simulated genome. Two replicate datasets

were generated for each category of rate heterogeneity. Other than that all simulations were run under a relaxed lognormal molecular clock (Drummond et al., 2006) with standard deviation set to 7.42×10^{-7}, HKY substitution model (Hasegawa et al., 1985) with the transition/transversion ratio parameter (κ) set to 6.0 and Γ-distributed rate heterogeneity with 4 categories and shape parameter 0.04 and empirical nucleotide frequencies, all derived from the results of the marginal likelihood analyses described earlier. A MERS-CoV sequence isolated from a camel (NRCE-HKU270) was provided as the starting state at the root. To include the effects of site-specific constraint we additionally carried out simulations under a Goldman-Yang codon model (Goldman and Yang, 1994) in πBUSS with empirical codon frequencies, κ=6.0 and dN/dS (ω) set to 0.1 (*i.e.* purifying selection) under same levels of rate heterogeneity and on the same phylogeny as the simulations described above. As these sequences were simulated on a tree of MERS-CoV, we refer to these datasets as being empirically simulated. We also reconstructed ancestral states for these sequences using ClonalFrameML, as described above, to arrive at a null expectation for the number of homoplasies we expect to observe under rate heterogeneity but without recombination.

Investigating the effects of temporal sampling and rate heterogeneity

All 10 sequence datasets simulated with fastsimcoal2 and 12 sequences empirically simulated in πBUSS were analyzed using LDhat (McVean et al., 2002) to ascertain the effects of temporal sampling, and in the case of πBUSS-simulated sequences, the effects of rate heterogeneity in the presence of absence of position-specific constraint. Additionally, empirically simulated sequence datasets were run through GARD (Kosakovsky Pond et al., 2006), since the method considers both differences in tree topology and branch lengths when calculating the likelihoods of trees. Stark rate heterogeneity amongst sites could thus easily be mis-interpreted as evidence for recombination by GARD, since trees derived from parts of the genome evolving at different rates will have different branch lengths and will fit the data better than a single tree.

Host-association alleles

In order to test for the presence of alleles associated with host shifts (presumably camel to human) we adapt the χ^2_{df} (Hedrick and Thomson, 1986) statistic of LD to estimate the association between host (camel or human) and alleles at polymorphic loci. Briefly, we consider the host to act as a polymorphic site (encoded as H or C, for human and camel, respectively) and compare the association between the "allele" or host and alleles at polymorphic sites. A perfect association of 1.0 could mean, for example, that a biallelic site has one allele that is only found in camel viruses and the other allele only in human viruses.

Results

MERS-CoV genome shows evidence of alternative tree topologies

GARD identified a breakpoint at nucleotide position 23722 (corrected ΔAIC=103.6 between single versus two tree model), roughly in the middle of the coding sequence for the S (spike) protein (see figure 3.1). The 2 phylogenies recovered from this breakpoint were incongruent (see figure A.1). Running the resulting fragment 1 (positions 1-23722) and fragment 2 (positions 23723-30126) through GARD again yielded a further breakpoint in fragment 1 at position 12257 (corrected ΔAIC=33.7), near the boundary between ORF1a and ORF1b genes. No more breakpoints could be identified by GARD in the resulting fragments 1.1 (positions 1-12257), 1.2 (positions 12258-23722) and 2 (positions 23723-30126).

5 out of 6 empirical simulation alignments under a nucleotide model simulated without recombination, were identified by GARD as having breakpoints around position 20000, where the clock rate for the rest of the "genome" was increased to be 1.3, 2 or 3 times higher than the first 20kb. Similarly, 2/6 alignments empirically simulated under a codon model were identified as having a breakpoint. Corrected ΔAIC values decreased with decreasing rate heterogeneity, indicating loss of statistical power to detect differences between genomic regions. Analyses in BEAST where the MERS-CoV genome is partitioned into positions 1-23722 and positions 23723-30126 (corresponding to the

first GARD-inferred breakpoint) with each partition having an independent molecular clock rate but the same tree or both independent molecular clock rates and independent trees, showed that rate heterogeneity as expressed by the ratio of second fragment rate to first fragment rate to be on the order of 1.513 (95% highest posterior density 1.275, 1.769) for unlinked clocks and 1.375 (95% HPDs 1.079, 1.707) for unlinked clocks and trees (see figure A.2). This result was derived from trees recovered for the model testing analyses described later (which are summarized in figure 3.5). As such, empirically simulated sequence data under 2-fold and 3-fold rate heterogeneity should be considered as caricature examples of a MERS-like organism, which we use to test the sensitivity of the methods we employ.

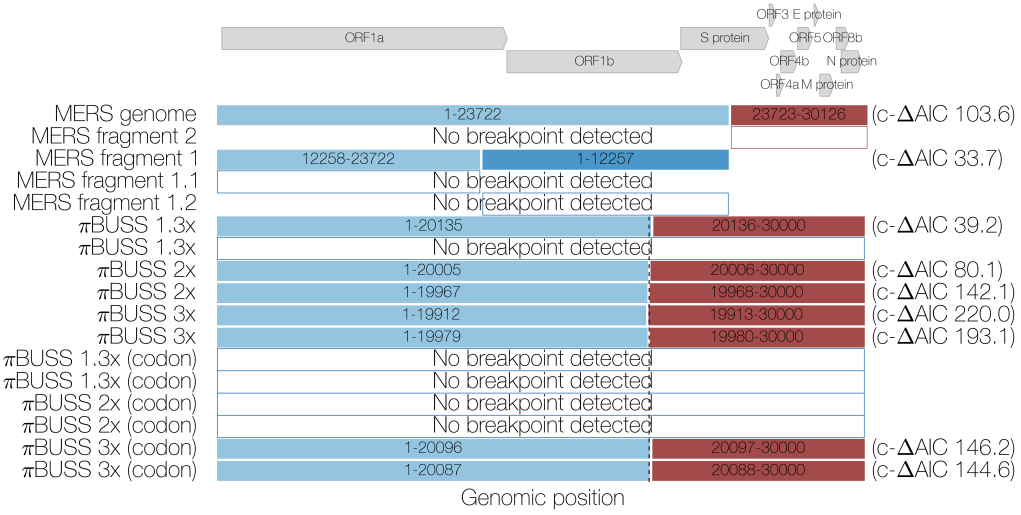


FIGURE 3.1: **Summary of GARD results.** Coloured boxes indicate fragments resulting from GARD-inferred breakpoints with corrected Δ-AIC values shown on the right. Dashed line indicates the actual position where the evolutionary model for simulated sequences under 3 levels of rate heterogeneity is changed. Arrows at the top indicate the positions and names of coding sequences within the MERS-CoV genome.

MERS-CoV genome exhibits linkage disequilibrium decay

Permutation tests as implemented in LDhat work under the assumption that loci are interchangeable only when there is free recombination or no recombination at all. The tests compare 4 statistics estimated from the actual data to 1000 permutations of the data where site numbers for each locus are reshuffled. Correlation coefficient between two measures of LD, r^2 (Hill and Robertson, 1968) and D' (Lewontin, 1964), are expected to show a negative correlation with increasing distance between loci if there

is recombination. Permutation of recombining loci will produce a distribution skewed towards more positive values for these two LD statistics and the percentile of the actual observed value can then be used to assess significance.

G4 is the sum of distances between pairs of loci with four observed haplotypes, which can only occur if there is repeat mutation or recombination at one of the loci. Under recombination the observed G4 statistic should take a statistically higher value in a distribution of G4 values derived from permuted data. Lkmax is the composite likelihood of pairs of loci under an estimated recombination rate and a given level of sequence diversity. Like the G4 statistic, this statistic is expected to fall in the upper tail of the distribution derived from permuted data in the presence of recombination.

All four permutation tests show a consistent signal of recombination in the MERS-CoV genome (see figures 3.2 and A.3). Data from fastsimcoal2 simulations, which did not have rate heterogeneity, produced values for these statistics which mostly fell inside the range of values generated by permuting the simulated data, as expected (figure 3.2). On one occassion this is not the case – simulation 9 passed the Lkmax test and failed the other three. Empirically simulated data, on the other hand, tended to exhibit extreme values, that is the observed value fell below the 2.5th or above the 97.5th percentile of the permuted data, but in ways which were not consistent with recombination. For example, replicate 1 of simulation with 3-fold rate heterogeneity under a nucleotide substitution model exhibits extreme values for all four tests, but only one of these – corr(r^2,d) is consistent with recombination.

Composite likelihood methods are susceptible to rate heterogeneity

The composite likelihood method, which finds the composite likelihood surface of recombination rate, inferred non-zero recombination rates for all simulated datasets (see figure A.4), revealing some degree of susceptibility to both temporal sampling and rate heterogeneity. A window-based approach of this test shows a sharp increase in the recombination rate estimated within 300 nucleotide windows around nucleotide 21000, close to the breakpoint inferred by GARD (see figure A.5). We recovered a qualitatively similar pattern when analyzing empirically simulated sequences (see figure A.5). It is important to note, however, that none of the simulated data, even under extreme

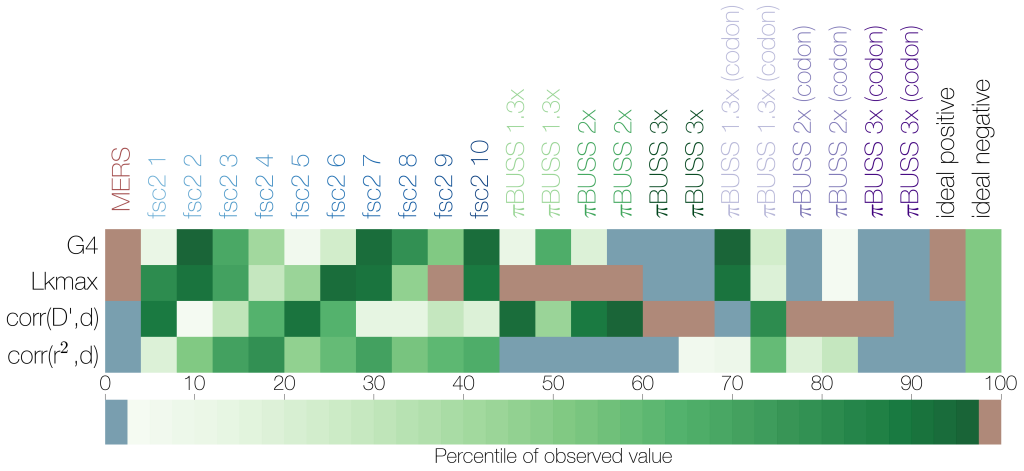


FIGURE 3.2: **Summary of non-parametric tests for recombination.** The percentile of the observed value for four statistics of LD decay (y axis) in the distribution of permuted datasets is indicated by colour. Sequence datasets are shown on the x axis, starting with MERS-CoV sequences, followed by 10 fastsimcoal2-simulated datasets and 12 empirically simulated datasets with different degrees of rate heterogeneity. Expected values for ideal datasets are shown in the last two columns, an ideal positive corresponds to the presence of recombination. Values falling between the 2.5th and 97.5th percentile are shown in green, values falling below the 2.5th percentile are in blue, those that are above the 97.5th percentile in red.

heterogeneity, reproduced the same scale of the estimated recombination rate. Whereas in MERS-CoV data the vast majority of 300 nucleotide windows after position 23000 have a recombination rate per base consistently higher than 0.005, only data simulated under extreme rate heterogeneity approach values as high as that.

In addition to the apparently higher recombination rate in regions with higher rates we expect rate heterogeneity to produce a higher density of polymorphic sites in regions that are evolving faster. This is quite obvious in empirically simulated data with 3-fold rate heterogeneity – the region with higher rate also contains, on average, more polymorphic loci per window in the last third of the "genome" than the first 20kb (see figure A.6). We only see hints of this in the actual MERS-CoV genome, with an apparent decline in polymorphism density from position 5000 to 15000 which resembles that of the simulated data with 1.3-fold rate heterogeneity.

Homoplasies in MERS-CoV genomes are ubiquitous

Homoplasy analyses suggest that the MERS-CoV genome is rife with apparent homoplasies. Both maximum likelihood and Bayesian approaches to ancestral sequence reconstruction converge on similar patterns of homoplasy density (figures 3.3 and A.7). Both methods identify the region around the gene encoding the S (spike) protein as having a high density of synonymous homoplasies.

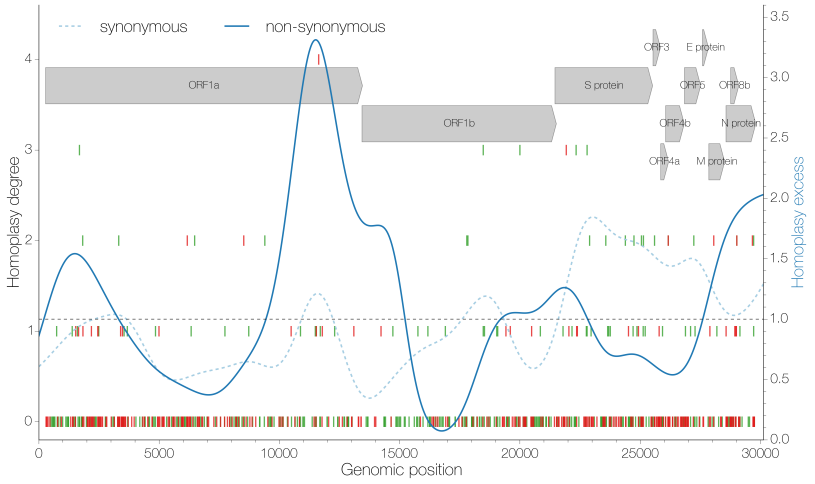


FIGURE 3.3: **Distribution of apparent homoplasies.** Position along the genome is shown on the x axis and homoplasy degree, the number of times a particular mutation has occured in excess in the tree as inferred by maximum likelihood, is shown on the y axis (left). Individual mutations are marked by vertical lines, synonymous ones in green and non-synonymous in red. The ratio of apparent homoplasy over synapomorphy kernel density estimates (bandwidth=0.1) is shown in blue for synonymous (dashed) and non-synonymous (solid) sites separately. Arrows at the top indicate the positions and names of coding sequences within the MERS-CoV genome.

Empirically simulated sequences showed that homoplasies are not that unlikely in the absence of recombination. All sequences empirically simulated in πBUSS under a nucleotide substitution model had 2-fold homoplasies ranging in frequency from 0.0222 to 0.0550 of all polymorphic sites, with sequences simulated under higher levels of rate heterogeneity having more homoplasies and higher homoplasy degrees (figure 3.4). However, even under a caricature model of rate heterogeneity we did not reach the same degree of homoplasy as that observed in MERS-CoV, where homoplasic sites comprise as much as 0.1447 of all polymorphic sites and reach homoplasy degrees as high as 4. We were able to recover similar proportions of homoplasic sites from empirical simulations under a codon model (ranging from 0.0821 to 0.1518 of all polymorphic sites),

but only under unrealistic values of rate heterogeneity. Even then, the distributions of homoplasy degrees for all simulated datasets are heavily skewed towards low homoplasy degrees, whereas the homoplasy degree distribution for MERS-CoV has a longer tail, indicating that homoplasies are more repeatable across the phylogeny.

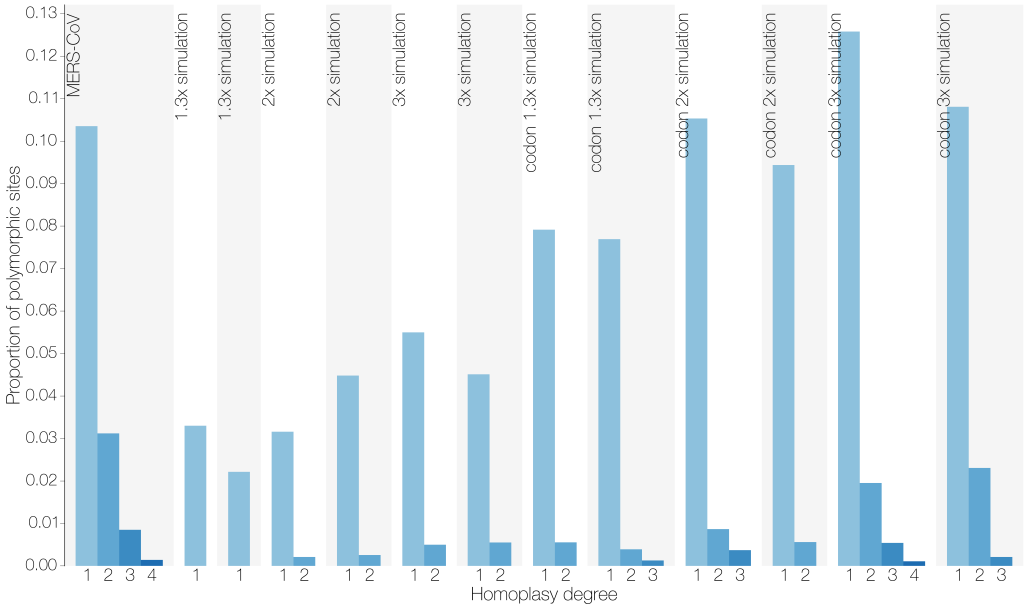


FIGURE 3.4: **Homoplasy prevalence in MERS-CoV and simulated datasets.** Bars show the proportion of all polymorphic sites that are homoplasic, split by homoplasy degree as inferred by maximum likelihood, in MERS-CoV and datasets simulated with different degrees of rate heterogeneity in the presence or absence of site-specific constraint in the form of a codon model. Homoplasy degree indicates how many times a given mutation has occured in excess in the phylogenetic tree.

Additional tests for excessive homoplasies, PHI and 3Seq also identify MERS-CoV sequences as being recombinant, albeit both spuriously identify some of the simulated datasets as recombinant (see figure A.10). Similarly to LDhat results, however, there were no cases where both methods falsely inferred the presence of recombination for the same dataset.

Model testing supports a model including rate heterogeneity, but not alternative tree topologies

A model including rate heterogeneity alone across breakpoints inferred by the GARD method (*i.e.* linked trees, unlinked relaxed clocks) performs best when applied to MERS-CoV data (figure 3.5 log marginal likelihoods: -48137.86 and -48138.91, using

path and stepping stone sampling, respectively). The next best-performing model (log Bayes factor ≈18) is linked trees and relaxed clocks. Overall, unlinking molecular clock rates between the two genomic partitions appears sufficient to dramatically improve model fit. Additionally, relaxed molecular clocks are preferred over strict molecular clocks (log BF>15, figure 3.5).

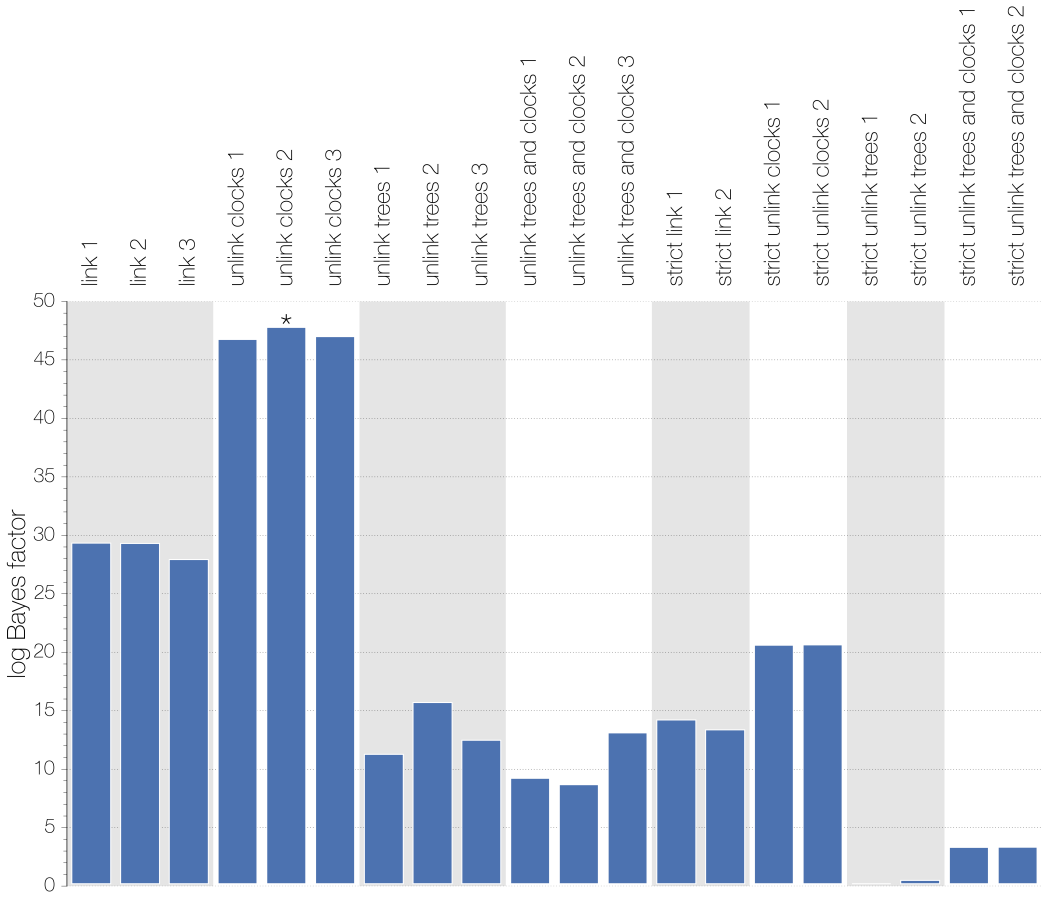


FIGURE 3.5: **Summary of model comparisons.** Difference in marginal likelihoods (Bayes factor) estimated by path-sampling between the worst model (linked strict molecular clock, unlinked trees) and all others. Asterisks indicate the best-performing model (unlinked relaxed clocks, linked trees, run 2) for MERS-CoV data. Analyses employing a relaxed molecular clock were run independently 3 times, those with a strict molecular clock 2 times. Marginal likelihoods estimated using stepping stone sampling gave identical results.

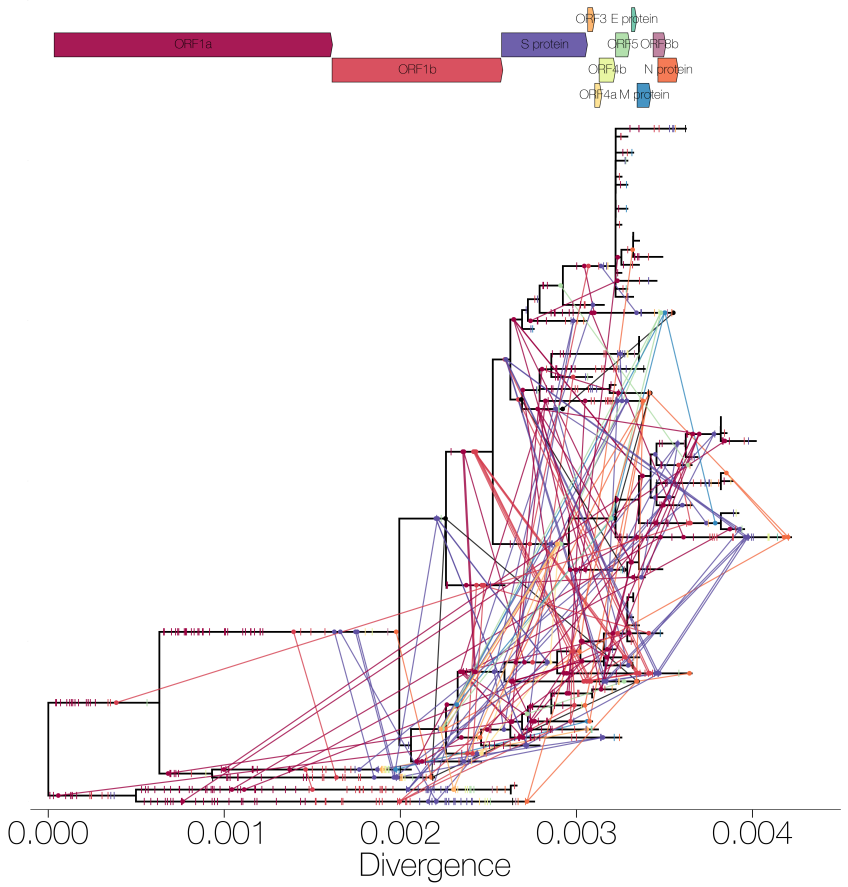


FIGURE 3.6: **Mutations mapped onto a ML phylogeny.** A maximum likelihood phylogeny of 85 MERS-CoV sequences with maximum likelihood-mapped mutations. Synapomorphies are shown as coloured ticks (coloured by coding sequence in which they occur) on branches where they occur. Homoplasies are shown as circles connected with coloured lines, colour corresponds with the coding sequence in which the mutation has occured. Mutations are positioned on the branches in proportion to where the mutation occurs in the genome, *e.g.* mutations shown towards the end of a branch correspond to mutations near the 3' terminus of the genome. Arrows at the top indicate the order, relative length and names of coding sequences within the MERS-CoV genome.

Discussion

Recombination tests consistently point to recombination in MERS-CoV

The majority of methods we used (with the exception of marginal likelihood model testing) point to the combined effects of recombination and rate heterogeneity in the genome

of MERS coronavirus. GARD (figure 3.1) identified 2 breakpoints in the genome with high support. It is important to note that calculating the likelihood of a phylogenetic tree involves the estimation of multiple parameters, which at the very least must include a particular tree topology and a set of branch lengths. As such, we considered that the inference of a breakpoint in the MERS-CoV genome could be caused by systematic rate heterogeneity, which we address by evaluating rate heterogeneity across the first breakpoint in BEAST. We estimate an empirical rate heterogeneity ratio between MERS-CoV genome positions 23723-30126 and 1-23722 to be on the order between 1.3 and 1.5 (see figure A.2). However, the support for this first breakpoint in MERS-CoV is comparable to support for empirically simulated sequences with 2-fold rate heterogeneity, and breakpoints under MERS-like levels of rate heterogeneity are difficult to detect. Overall, this suggests that evidence for differences in likelihoods between MERS fragments 1 and 2 are beyond what would be expected from rate heterogeneity alone.

Permutation tests show that statistics related to LD decay derived from MERS-CoV sequence data are outliers compared to permuted data (figures 3.2 and A.3). Sequences simulated empirically with varying levels of rate heterogeneity, under nucleotide or codon models of substitution, also have a tendency to exhibit extreme values for these statistics. However, only MERS-CoV data has values for all 4 tests that are in the direction consistent with recombination. For simulated datasets, especially those simulated under extreme levels of rate heterogeneity, values deviating significantly from the permuted data were recovered, but often indicated internally contradictory scenarios.

Homoplasy analyses regardless of inference method show that MERS-CoV sequences contain a large number of homoplasic sites with high homoplasy degrees (figures 3.3 and A.7). Through sequence simulation we also confirmed that both the numbers of homoplasic sites and their homoplasy degrees in MERS-CoV genomes are excessive, even when compared to unrealistic scenarios (*e.g.* 3-fold rate heterogeneity, see figure 3.4). Homoplasies become increasingly more prevalent when a more realistic codon model is used, due to differences in codon position constraint. Even then, MERS-CoV genomes possess mutations with much higher homoplasy degrees, surpassing simulated datasets with caricature levels of rate heterogeneity. This makes sense under a recombination scenario, as alleles persist in a diverse population and get recombined into novel backgrounds repeatedly, giving an appearance of highly repeatable mutations. Nevertheless,

substitution patterns in real genomes are often highly complex and homoplasy-based methods have been shown to be susceptible to rate heterogeneity across sites, especially under higher levels of sequence divergence (Posada and Crandall, 2001). Although rate heterogeneity certainly exists in MERS-CoV data, the divergence levels are still quite low (θ/site = 0.0047), giving us some degree of certainty in our inference of homoplasies. It is also reassuring that both maximum likelihood and Bayesian sequence reconstruction converged on similar patterns of homoplasy and synapomorphy across the genome (figures 3.3 and A.7). This is important, since homoplasies inferred using BEAST are integrated over all possible tree topologies, whereas homoplasies inferred by maximum likelihood were conditioned on a single tree. The convergence between these two methods suggests that the data contain enough phylogenetic signal to recover what could be called a "true" tree and that homoplasies, for the most part, can be correctly identified as such.

One major concern surrounding the inference of homoplasies is parallel evolution, *e.g.* host adaptation. There are a number of canonical mutations associated with host shifts, *e.g.* the glutamic acid to lysine amino acid substitution at position 627 in the PB2 protein of avian influenza A viruses confers the ability of the virus to replicate in mammals (Subbarao et al., 1993) and a small number of amino acid substitutions in Parvoviruses are associated with adaptations to different hosts (Chang et al., 1992). If MERS-CoV is repeatedly emerging in humans convergent mutations would be expected to arise that might allow the virus to adapt to humans.

However, we expect most host-adaptation mutations to be non-synonymous, whereas we detect both non-synonymous and synonymous homoplasies. Although synonymous sites can be under selection, indeed this has been shown for influenza A viruses (Gog et al., 2007), the function of synonymous sites is not easy to assess using computational methods alone. Thus, if we assume that synonymous sites in the genome of MERS-CoV are not undergoing convergent evolution then only the action of recombination, rather than repeated selection for the same host-specific mutations, can explain the data better. Furthermore, we do not detect any strong associations between host (human or camel) and particular alleles (figure A.8), although we do not believe that there is a sufficient number of sequences from camels to have much confidence in this result.

The overall phylogenetic and genomic patterns of homoplasies are consistent with fairly frequent recombination through time (figure 3.6). Recent recombination should result in long homoplasy tracts shared across branches in the phylogenetic tree. At most we observe 2 stretches of adjacent homoplasies, one encompassing 3 homoplasies and another encompassing 2 homoplasies, that are shared between taxa, and likely to be caused by recent recombination. The vast majority of homoplasies that we observe, however, occur on their own. Recombination tracts, rather than single template switches are not uncommon in other coronaviruses (Keck et al., 1988; Kottier et al., 1995; Herrewegh et al., 1998). Thus in MERS-CoV we interpret extremely short homoplasy tracts as evidence of relatively frequent recombination. Alternatively, recombination tracts might be short and thus unable to transfer multiple informative sites across lineages.

Unlike all other tests we performed model testing through marginal likelihoods indicates that models including rate heterogeneity explain MERS-CoV data partitioned across a well-supported breakpoint better than models including independent trees. At first this may seem paradoxical, but we believe this result is due to the combined effects of the way homoplasious sites are distributed across the genome and phylogenetic tree of MERS-CoV (figure 3.6) and the number of parameters involved. A speckled pattern of homoplasious sites without phylogenetic signal could easily be overwhelmed by the signal coming from the sites that support what could be called "the one true tree", *i.e.* the clonal frame, in the data. Secondly, each phylogenetic tree contains at least n-1 free parameters, so it is not surprising then that models attempting to recover 2 independent trees for both genomic fragments resulting from alternative tree topology analysis of MERS-CoV with highly correlated genealogies are penalized for the extra parameters introduced by a second tree. We would additionally like to point out that the fit of models including relaxed molecular clocks result in dramatic improvements to model fit compared to models with strict molecular clocks (log BF>15, figure 3.5). Although this could be interpreted as evidence for a considerable degree of lineage rate heterogeneity, the more parsimonious explanation is the ability of a relaxed molecular clock to accommodate homoplasies of recombinant origin, which do not necessarily accumulate at a relatively constant rate like genuine *de novo* mutations do.

Implications for future analyses

Recombination aside, MERS-CoV genomes exhibit a significant degree of rate heterogeneity amongst sites. Marginal likelihood analyses indicate that estimating independent molecular clocks after partitioning the MERS-CoV genome into two fragments alone substantially increases model fit over a completely linked (trees and clocks) model (log Bayes factor ≈18). This highlights the advantage of employing relaxed molecular clocks, as in our case the method is clearly capable of accomodating recombination in an otherwise entirely clonal analysis framework. Recombination transfers alleles acquired via mutations in one genetic background to another genetic background, which under the correct reconstruction of the phylogeny will appear as a burst of *de novo* mutations on the branch receiving the recombination tract, which will appear as strong amongst-lineage rate heterogeneity. This is exactly the problem that relaxed molecular clocks have been designed to tackle. In addition, previous studies of SARS-like coronaviruses in bats have identified recombination breakpoints in small numbers of isolates falling close to the "transition zone" around site 22000 (Hon et al., 2008; Lau et al., 2010) which in our analysis of MERS-CoV is where GARD, LDhat and BEAST identify changes in the underlying model of evolution (figures 3.1, A.5 and 3.3). Overall, a more detailed investigation will need to be done to determine if empirical patterns of rate variation in MERS-CoV have the potential to generate apparent recombination signals.

We also expect that as more sequences of MERS-CoV become available more homoplasies will be detected, some contributing to the homoplasy degree of the homoplasies already reported here, some previously unknown and some turning mutations currently thought of as synapomorphies into homoplasies. Although new sequences are likely to come from human cases, we think that sequencing MERS-CoV circulating in dromedary camels is of extreme importance from both surveillance and epidemiological points of view. Understanding the prevalence of the virus in camels, as well as knowing where, when and how often zoonotic events occur would advance our understanding of MERS-CoV epidemiology far beyond what can be learned from human cases alone.

Implications about the virus population structure and infection dynamics

Our results point towards frequent recombination in MERS-CoV in the recent history of the MERS-CoV outbreak. For this to occur different lineages of the virus must encounter each other often and implies frequent co-infection with MERS-CoV. To date it is difficult to ascertain whether the human infections with MERS-CoV are a result of substantial asymptomatic transmission amongst humans, or repeated zoonosis of the virus from camels to humans or a combination thereof. Given the severity of MERS we find it unlikely that humans could be sufficiently frequently co-infected with two or more different lineages of the virus. Previous serological studies have failed to find evidence of prevalent past MERS-CoV infections of humans (Gierer et al., 2013; Aburizaiza et al., 2013), although a recent nation-wide study in Saudi Arabia has detected non-negligible numbers of individuals with antibodies against MERS-CoV, especially amongst shepherds and slaughterhouse workers (Müller et al., 2015). We thus propose that MERS-CoV mostly infects, and recombines, in camels. A study by Adney et al. (2014) has shown that camels only suffer mild symptoms from MERS-CoV infection and numerous other studies indicate an extremely high prevalence of antibodies specific against MERS-CoV in camels (Müller et al., 2014; Corman et al., 2014b; Chu et al., 2014; Reusken et al., 2013, 2014). At the same time, however, sequencing has not indicated the presence of multiple infection in camels, or any other animal. We believe that individual MERS-CoV co-infections are rare, but given the size of the epidemic in camels, as inferred from serology, the total number of co-infections is high. In addition, MERS-CoV infection is transient in camels Adney et al. (2014) and thus sequencing efforts, which have been insufficient and very limited in camels, are highly unlikely to capture a co-infection.

Another point worth considering is that alleles that have arisen through mutation in MERS-CoV can be recombined, increasing the genetic variation of the virus (Muller, 1932). Whether this is of epidemiological importance for humans depends entirely on what alleles are circulating in the reservoir, although there is no evidence that MERS-CoV is particularly likely to become as transmissible as common human pathogens or even SARS-CoV.

Data availability

Python scripts used to process trees and sequences are available at:

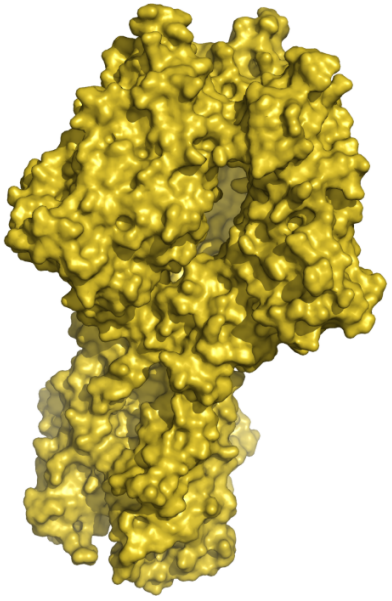
https://github.com/evogytis/MERS_recombination/tree/master/scripts.

Input and output files for programs used are publicly available at:

https://github.com/evogytis/MERS_recombination.

CHAPTER 4

REASSORTMENT BETWEEN INFLUENZA B LINEAGES AND THE EMERGENCE OF A CO-ADAPTED PB1-PB2-HA GENE COMPLEX



Deep in the human unconscious is a pervasive need
for a logical universe that makes sense.
But the real universe is always one step beyond logic.

– Frank Herbert, *Dune*

This chapter was published as **Dudas G, Bedford T, Lycett S, Rambaut A. 2015**. Reassortment between influenza B lineages and the emergence of a coadapted PB1–PB2–HA gene complex. *Molecular Biology and Evolution* **32**:162–172. 10.1093/molbev/msu287. under the same title. I analyzed the data and wrote the manuscript. T. Bedford and A. Rambaut provided editorial assistance. S. Lycett provided the normalization procedure for tree statistic comparisons.

Introduction

Seasonal influenza causes between 250,000 and 500,000 deaths annually and comprises lineages from three virus types (A, B and C) co-circulating in humans, of which influenza A is considered to cause the majority of seasonal morbidity and mortality (World Health Organization, 2009). Occasionally influenza B viruses become the predominant circulating virus in some locations, for example in the 2012/2013 European season as many as 53% of influenza sentinel surveillance samples tested positive for influenza B (Broberg et al., 2013).

Like other members of *Orthomyxoviridae*, influenza B viruses have segmented genomes, which allow viruses co-infecting the same cell to exchange segments, a process known as reassortment. Influenza A viruses are widely considered to be a major threat to human health worldwide due to their ability to cause pandemics in humans via reassortment of circulating human strains with non-human influenza A strains. Although influenza B viruses have been observed to infect seals (Osterhaus et al., 2000; Bodewes et al., 2013) through an anthroponosis, they are thought to primarily infect humans and are thus unlikely to exhibit pandemics due to the absence of an animal reservoir from which to acquire antigenic novelty. Both influenza A and B evolve antigenically through time in a process known as antigenic drift, in which mutations to the haemagglutinin (HA) protein allow viruses to escape existing human immunity and persist in the human population, leading to recurrent seasonal epidemics (Burnet, 1955; Hay et al., 2001; Bedford et al., 2014).

Currently circulating influenza B viruses comprise two distinct lineages – Victoria and Yamagata (referred to as Vic and Yam, respectively) – named after strains B/Victoria/2/87 and B/Yamagata/16/88, that are thought to have genetically diverged in HA around 1983 (Rota et al., 1990). These two lineages now possess antigenically distinct HA surface glycoproteins (Kanegae et al., 1990; Rota et al., 1990; Nerome et al., 1998; Nakagawa et al., 2002; Ansaldi et al., 2003) allowing them to co-circulate in the human population. Phylogenetic analysis of evolutionary rate, selective pressures and reassortment history of influenza B has shown extensive and often complicated patterns of reassortment between all segments of influenza B viruses both between and within the Vic and Yam lineages (Chen and Holmes, 2008).

Here, we extend previous methods to reveal an intriguing pattern of reassortment in influenza B. In our approach, membership to either the Victoria or Yamagata lineage in one segment is used to label the individual isolates in the tree of the other segments. By modelling the transition between labels on a phylogenetic tree, reassortment events which result in the replacement of one segment's lineage by another show up as label changes along a branch (Figure 4.1). We use this method to reconstruct major reassortment events and quantify reassortment dynamics over time in a dataset of 452 influenza B genomes, and conduct secondary analyses in a dataset of 1603 influenza B genomes.

We show that despite extensive reassortment, three of the eight segments – two segments coding for components of the influenza B virus polymerase, PB1 and PB2, and the surface glycoprotein HA – still survive as distinct Victoria and Yamagata lineages, which appear to be co-dependent to the point where virions which do not contain PB1, PB2 or HA segments derived entirely from either the Vic or the Yam lineage have rarely been isolated and only circulate as transient lineages once isolated. In other segments (PA, NP, NA, MP and NS) a single lineage has introgressed into the opposing background and replaced the previous lineage: all currently circulating influenza B viruses have PA, NP, NA and MP segments derived from Yamagata lineage and NS segments derived from Victoria lineage.

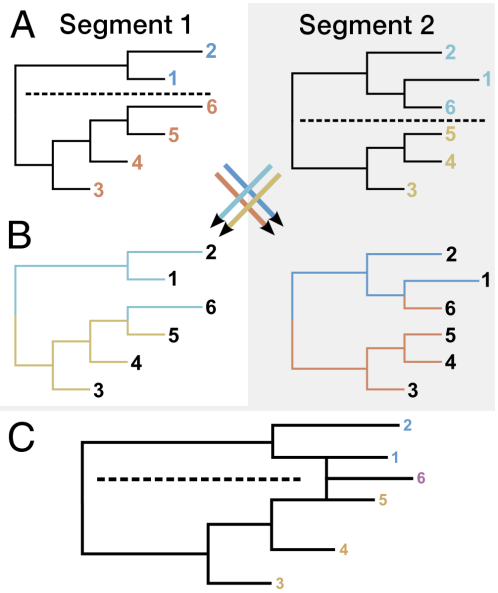


FIGURE 4.1: **Schematic analysis of reassortment patterns.** A) We begin by assigning sequences falling on either side of a specified bifurcation within each segment tree to different lineages, in this case, the Victoria and Yamagata bifurcation that occurred in the early 1980s. B) We then transfer lineage labels from one tree to the same tips in another tree. Transitions between labels along this second tree thus indicate reassortment events that combine lineages falling on different sides of the Vic/Yam bifurcation in the first tree. C) A reassortment graph depiction shows that tip number 6 is determined to be a reassortant based on B).

Methods

We compiled a primary dataset of 452 complete influenza B genomes from GISAID (Bogner et al., 2006) dating from 1984 to 2012. The longest protein coding region of each segment was extracted and used for all further analyses. We thus assume that homologous recombination has not taken place and that the evolutionary history of the whole segment can be inferred from the longest coding sequence in the segment. To date there has been little evidence of homologous recombination in influenza viruses (Chare et al., 2003; Boni et al., 2008; Han et al., 2010). The segments of each strain were assigned to either Vic or Yam lineage by making maximum likelihood trees of each segment using PhyML (Guindon and Gascuel, 2003) and identifying whether the isolate was more closely related to B/Victoria/2/87 or B/Yamagata/16/88 sequences in that segment, with the exception of the NS segment since B/Victoria/2/87 was a reassortant and possessed a Yam lineage NS (Lindstrom et al., 1999). B/Czechoslovakia/69/1990 was considered as being representative of Victoria lineage for the NS segment. Every segment in each genome thus received either a Vic or a Yam lineage designation, for

example the strain B/Victoria/2/87 received V-V-V-V-V-V-V-Y, since its NS segment is derived from the Yam lineage and the rest of the genome is Vic.

We also collated a secondary dataset from all complete influenza B virus genomes available on GenBank as of May 5, 2014. After removing isolates that had considerable portions of any sequence missing, were isolated prior to 1980 or were suspected of having a contaminant sequence in any segment, we were left with 1603 sequences. This dataset only became available after all primary analyses were performed, is mainly composed of sequences from Australia, New Zealand and the United States and are too numerous to analyze in BEAST (Drummond et al., 2012). PhyML (Guindon and Gascuel, 2003) was used to produce phylogenies of each segment and the lineage of each isolate was determined based on grouping with either B/Victoria/2/87 or B/Yamagata/16/88 sequences, as described above. By associating strains with lineage identity of each of their segments, we reconstructed the most parsimonious inter-lineage reassortment history for the secondary dataset. The secondary dataset was used to check how representative the primary dataset was, to estimate linkage disequilibrium (LD) and to broadly confirm our results. All analyses pertain to the primary dataset unless stated otherwise.

Temporally-calibrated phylogenies were recovered for each segment in the primary dataset using Markov chain Monte Carlo (MCMC) methods in the BEAST software package (Drummond et al., 2012). We modeled the substitution process using the HKY model of nucleotide substitution (Hasegawa et al., 1985), with separate transition models for each of the 3 codon partitions, and additionally estimated realized synonymous and non-synonymous substitution counts (O'Brien et al., 2009). We used a flexible Bayesian skyride coalescent model (Minin et al., 2008). We accounted for incomplete sampling dates for 94 sequences (of which 93 had only year and one had only year and month of isolation) whereby tip date is estimated as a latent variable in the MCMC integration. A relaxed molecular clock was used, where branch rates are drawn from a lognormal distribution (Drummond et al., 2006). We ran three independent MCMC chains, each with 200 million states, sampled every 20,000 steps and discarded the first 10% of the MCMC states as burn-in. After assessing convergence of all three MCMC chains by visual inspection using Tracer (Rambaut et al., 2009), we combined samples across chains to give a total of 27,000 samples from the posterior distribution of trees.

Every sequence was assigned seven discrete traits in BEAUti corresponding to the lineages of all other segments with which a strain was isolated *e.g.* PB1 tree had PB2, PA, HA, NP, NA, MP and NS as traits and V or Y as values for each trait. We inferred the ancestral state of lineages in each segment by modelling transitions between these discrete states using an asymmetric transition matrix (Lemey et al., 2009b) with Bayesian stochastic search variable selection (BSSVS) to estimate significant rates. Because the posterior set of trees for a single segment has branches labelled with the inferred lineage in the remaining seven segments, we can detect inter-lineage reassortments between pairs of segments by observing state transitions, i.e. Yam to Vic or Vic to Yam (Figure 4.1). In addition, by reconstructing the ancestral state of all other genomic segments jointly we can infer co-reassortment events when more than one trait transition occurs on the same branch in a tree. Inter-phylogeny labeling approaches have been extensively used in the past to investigate reticulate evolution in influenza A viruses and HIV (Lycett et al., 2012; Ward et al., 2013; Lu et al., 2014).

Measures of diversity

We inferred the diversity of each segment from their phylogenetic tree by estimating the date of the most recent common ancestor of all branches at yearly time points, which places an upper bound on the maximum amount of diversity existing at each time point. A version of this lineage turnover metric has previously been used to investigate the tempo and strength of selection in influenza A viruses during seasonal circulation (Rambaut et al., 2008; Bedford et al., 2011). In addition, we calculated mean pairwise time of most recent common ancestor (TMRCA) between branches labelled as Vic and Yam for PB1, PB2 and HA traits. This gave us a measure of how much a particular segment reassorts with respect to Vic and Yam lineages of PB1, PB2 and HA segments. If Vic and Yam lineages of PB1, PB2 and HA segments were to be considered as being separate populations this measure would be equivalent to 'between population' diversity.

We also calculated the total amount of sampled evolutionary time spent by each segment with entirely Vic, entirely Yam or mixed lineage PB1, PB2 and HA segments. We do this by summing the branch lengths in each tree under 3 different lineage combinations of the PB1, PB2 and HA segments: PB1-PB2-HA derived entirely from Yamagata

lineage, PB1-PB2-HA entirely derived from Victoria lineage and PB1-PB2-HA derived from a mixture of the two lineages. This gives a measure of how successful, over long periods of time, each particular PB1-PB2-HA constellation has been.

Tree to tree similarities

We express the normalized distance Δ_{TMRCA} between trees belonging to two segments A and B for a particular posterior sample i, following

$$\Delta_{\text{TMRCA}}(A_i, B_i) = \frac{\delta_{\text{TMRCA}}(A_i, A_i') + \delta_{\text{TMRCA}}(B_i, B_i')}{2\,\delta_{\text{TMRCA}}(A_i, B_i)}, \tag{4.1}$$

where $\delta_{\text{TMRCA}}(A_i, B_i) = \frac{1}{n}\sum_{j=1}^{n} g(A_{ij}, B_{ij})$ and n is the total number of pairwise comparisons available between sets of tips. Thus, $g(A_{ij}, B_{ij})$ is the absolute difference in TMRCA of a pair of tips j, where the pair is drawn from the ith posterior sample of tree A and the ith posterior sample of tree B. Additionally, $\delta_{\text{TMRCA}}(A_i, A_i')$ is calculated from the ith posterior sample of tree A and ith posterior sample of an independent analysis of tree A (which we refer to as A'), which is used in the normalization procedure to control for variability in tree topology stability over the course of the MCMC chain (see Figures B.15 and B.16). We had 3 replicate analyses of each segment and in order to calculate $\delta_{\text{TMRCA}}(A_i, A_i')$ we used analyses numbered 1, 2 and 3 as A and analyses numbered 2, 3 and 1 as A', in that order. We subsampled our combined posterior distribution of trees to give a total of 2,700 trees on which to analyze Δ_{TMRCA}.

Calculating the normalized $\Delta_{\text{TMRCA}}(A_i, B_i)$ for each MCMC state provides us with a posterior distribution of this statistic allowing specific hypotheses regarding similarities between the trees of different segments to be tested. Our approach exploits the branch scaling used by BEAST (Drummond et al., 2012), since the trees are scaled in absolute time and insensitive to variation in nucleotide substitution rates between segments, allowing for direct comparisons between TMRCAs in different trees. In the absence of reassortment we expect the tree of every segment to recapitulate the 'virus tree', a concept analogous to 'species trees' in population genetics. Our method operates under the assumption that the segment trees capture this 'virus tree' of influenza B viruses quite well. It is not an unreasonable assumption, given the seasonal bottlenecks influenza viruses experience. This makes it almost certain that influenza viruses circulating at

any given time point are derived from a single genome that existed in the recent past. The δ_{TMRCA} statistic essentially quantifies the temporal distance between admixture events and nodes in the 'virus tree' (see Figure B.17). We normalize δ_{TMRCA} values to get Δ_{TMRCA}, a measure which quantifies the extent to which the similarity of two independent trees resembles phylogenetic noise. The δ_{TMRCA} statistic is an extension of patristic distance methods and has previously been used to tackle a wide variety of problems, as phylogenetic distance in predicting viral titer in *Drosophila* infected with viruses from closely related species (Longdon et al., 2011) and to assess temporal incongruence in a phylogenetic tree of amphibian species induced by using different calibrations (Ruane et al., 2011).

Linkage disequilibrium across the influenza B genome

We used the secondary GenBank dataset with 1603 complete genome sequences to estimate linkage disequilibrium (LD) between amino acid loci across the longest proteins encoded by each segment of the influenza B virus genome. To quantify LD we adapt the χ^2_{df} statistic from (Hedrick and Thomson, 1986):

$$\chi^2_{df} = \frac{\chi^2}{N\,(k-1)\,(m-1)}, \tag{4.2}$$

where χ^2 is calculated from a classical contingency table, N is the number of haplotypes and $(k-1)(m-1)$ are the degrees of freedom. This statistic is equal to the widely used r^2 LD statistic at biallelic loci, but also quantifies LD when there are more than two alleles per locus (Zhao et al., 2005). LD was estimated only at loci where each nucleotide or amino acid allele was present in at least two isolates. We ignored gaps in the alignment and did not consider them as polymorphisms. In all cases we used a minor allele frequency cutoff of 1%. We also calculated another LD statistic, D' (Lewontin, 1964) as $D'_{ij} = D_{ij}/D^{max}_{ij}$, where $D_{ij} = p(A_iB_j) - p(A_i)p(B_j)$ and

$$\begin{aligned} D^{max}_{ij} &= min[p(A_i)p(B_j), (1-p(A_i))(1-p(B_j))] \text{ when } D_{ij} < 0 \\ D^{max}_{ij} &= min[(1-p(A_i))p(B_j), p(A_i)(1-p(B_j))] \text{ when } D_{ij} \geq 0, \end{aligned} \tag{4.3}$$

where $p(A_i)$ is the frequency of allele A_i at locus A, $p(B_j)$ is the frequency of allele B_j at locus B and $p(A_iB_j)$ is the frequency of haplotype A_iB_j. D' is inflated when some haplotypes are not observed *e.g.* when the minor allele frequency is low. We find that D' is almost uniformly high across the influenza B virus genome and close to 1.0 for almost any pair of polymorphic loci. This is because most amino acid alleles in the population exist transiently, meaning that they do not get a chance to reassort and we only observe them within the backgrounds of more persistent alleles, which D' quantifies as complete LD. We think that metrics related to r^2, like χ^2_{df}, perform much better on temporal data such as ours in finding persistent associations between alleles and are easier to interpret.

Data availability

Python scripts used to process trees and sequences are available at:
https://github.com/evogytis/fluB/tree/master/scripts.

Output files from scripts, lineage designations, MCC trees, acknowledgment tables, accession numbers and redacted XML files (per GISAID Data Access Agreement) are publicly available at:
https://github.com/evogytis/fluB/tree/master/data.

Results

Analysis of reassortment patterns across Victoria and Yamagata lineages

The differentiation into Vic and Yam lineages can be seen in all segments (Figure 4.2) and is followed by inter-lineage reassortment events. In the phylogenetic trees of the PA, NP, NA, MP and NS segments either the Victoria or Yamagata lineage has become the 'trunk' of the tree, with present-day viruses deriving entirely from the Victoria or Yamagata lineage (yellow vs purple bars to the right of each tree in Figure 4.2) following reassortment. However, the Victoria and Yamagata lineages of PB1, PB2 and HA segments continue to co-circulate to this day. Periodic loss of diversity in PA,

NP, NA, MP and NS segments is consistent with introgression of one lineage into the other in those segments, while maintenance of parallel Victoria and Yamagata lineages results in continually increasing diversity in segments PB1, PB2 and HA (Figure 4.3). The PB1, PB2 and HA segments from present-day viruses each maintain a common ancestor in ∼1983 and thus accumulate genetic diversity since the split of those segments into Vic and Yam lineages, while other segments often lose diversity with ancestors to present-day viruses appearing between ∼1991 and ∼1999.

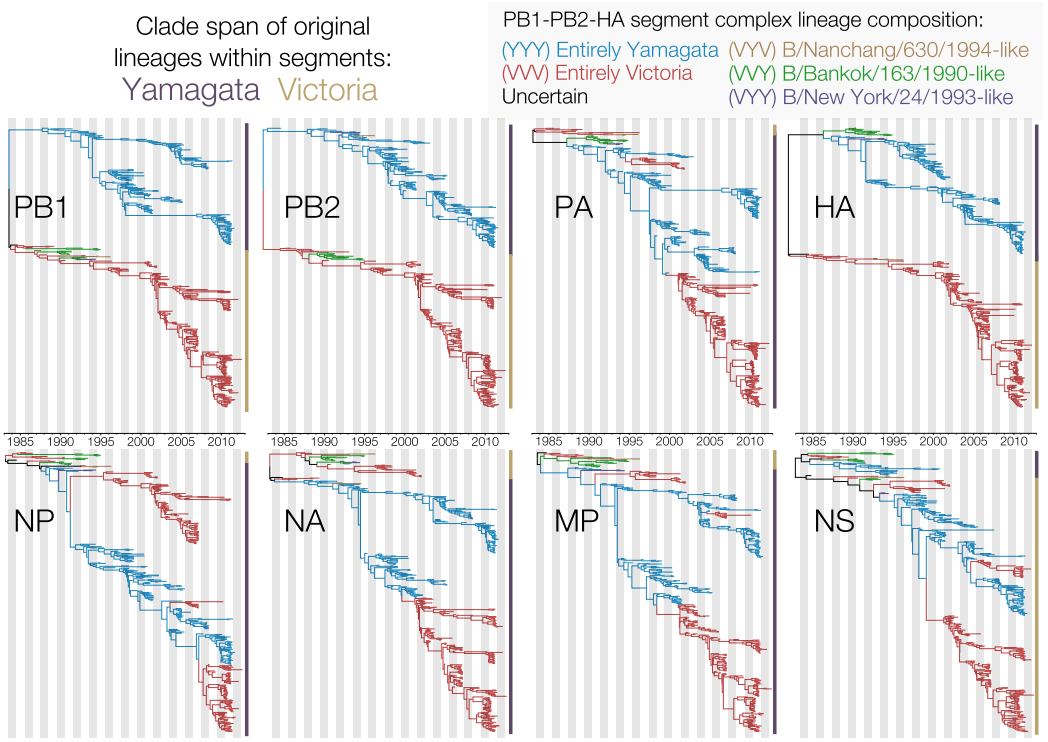


FIGURE 4.2: **Maximum clade credibility (MCC) trees of all 8 genome segments of influenza B viruses isolated since 1980.** Trees are coloured based on inferred PB1-PB2-HA lineage. Vertical bars to the right of each tree indicate the original Victoria and Yamagata lineages within each segment. Each tree is the summarised output of a single analysis comprised of 9000 trees sampled from the posterior distribution of trees.

By measuring mean pairwise diversity between branches in each tree that were assigned either a Vic or Yam label in other segments, we look for reductions in between-lineage diversity, which indicate that an inter-lineage reassortment event has taken place (Figure 4.4). This method gives a quantitative measure of reassortment-induced loss of diversity between Victoria and Yamagata lineages in two trees, although care should be taken when interpreting the statistic, as it does not correspond to any real TMRCAs in the tree, but can be interpreted as mean coalescence date between Vic and Yam

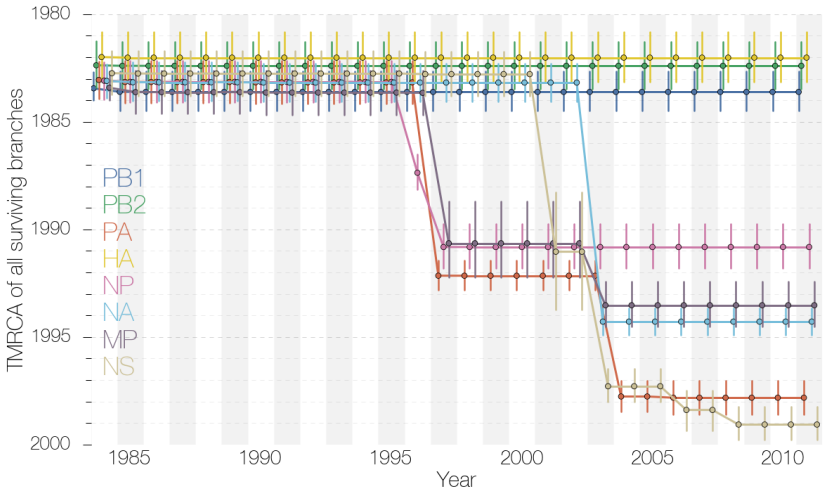


FIGURE 4.3: **Oldest TMRCA of all surviving branches over time.** PA, NP, NA, MP and NS segments of influenza B viruses show periodic increases in TMRCAs of all surviving branches (indicative of diversity loss), suggesting lineage turnover. PB1, PB2 and HA segments, on the other hand, maintain the diversity dating back to the initial split of Vic and Yam lineages. Each point is the mean time of most recent common ancestor (TMRCA) of all surviving lineages existing at each time slice through the tree and vertical lines indicating uncertainty are 95% highest posterior densities (HPDs).

lineages of PB1, PB2 and HA segments in all other trees. We focus only on PB1, PB2 and HA lineage labels, since all other segments eventually become completely derived from either the Vic or the Yam lineage. Losses of diversity (represented by more recent mean pairwise TMRCAs between Vic and Yam labels) in Figure 4.4 indicate that every segment has reassorted with respect to the Victoria and Yamagata lineages of PB1, PB2 and HA segments. However, we also see that the labels for these three segments show reciprocal preservation of diversity after 1997. This suggests that after 1997 no reassortment events have taken place between Victoria and Yamagata lineages of PB1, PB2 and HA segments and their lineage labels only 'meet' at the root. We do see reduced diversity between Vic and Yam labels of PB1, PB2 and HA segments in a time period close to the initial split of Vic and Yam lineages (1986–1996). These reductions in diversity represent small clades with reassortant PB1-PB2-HA constellations, which go extinct by 1997 (see Figure 4.2). We also observe that the assignment of these three segment labels to branches of other segment trees is very similar and often identical after 1997. This suggests that PB1, PB2 and HA lineage labels switch simultaneously in all trees after 1997.

We show the ratio of Vic to Yam sequences in our primary and secondary datasets in

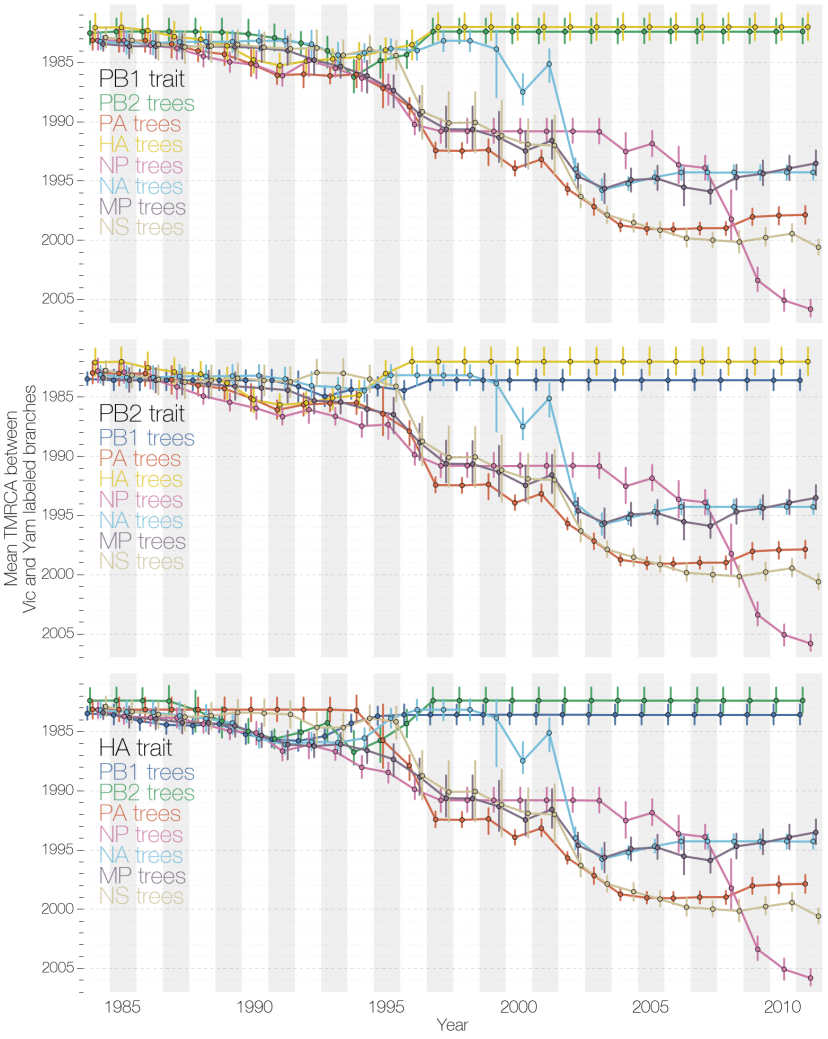


FIGURE 4.4: **Mean pairwise TMRCA between Vic and Yam branches under PB1, PB2 and HA label sets.** PB1, PB2 and HA segment labels indicate that these segments show reciprocal preservation of diversity, which dates back to the split of Vic and Yam lineages. All other segments show increasingly more recent TMRCAs between branches labelled as Vic and Yam in PB1, PB2 and HA label sets. All vertical lines indicating uncertainty are 95% highest posterior densities (HPDs).

different influenza seasons in Figure 4.5, which is based on which lineage each sequence was assigned to (see Methods). It is evident that losses of diversity in the PA, NP, NA, MP and NS segments are related to either the Vic (NS) or the Yam (PA, NP, NA, MP) lineage replacing the other lineage in the influenza B virus population. Similarly, the lack of reassortment between Vic and Yam lineages and maintenance of diversity of PB1, PB2 and HA can be seen, where the two lineages have been sequenced at a ratio close to 50% over long periods of time (Figure 4.5). On a year-to-year basis, however, the ratios for Vic and Yam sequences PB1, PB2 and HA can fluctuate dramatically consistent with one lineage predominating within a given season, in agreement with

surveillance data (Reed et al., 2012).

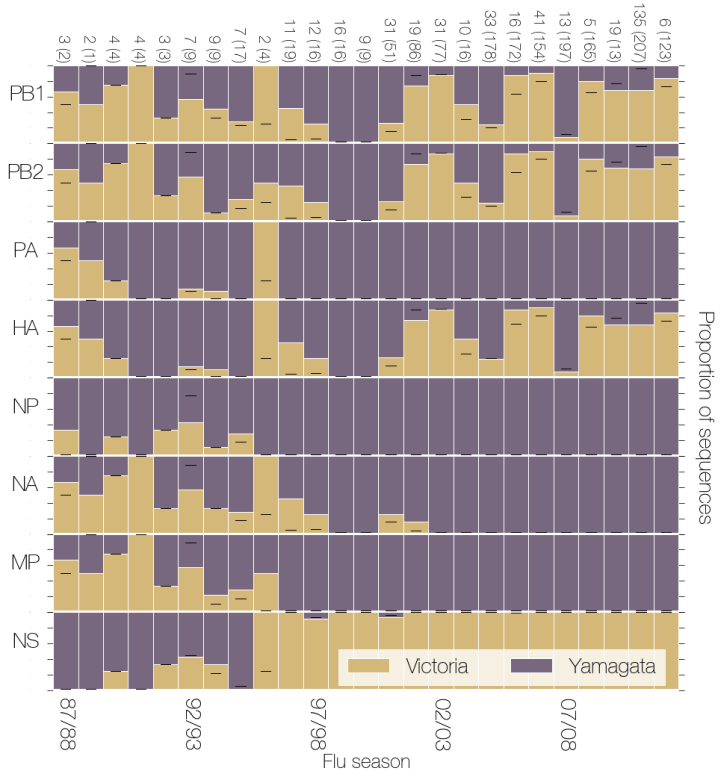


FIGURE 4.5: **Ratio of Vic and Yam sequences in the dataset.** The ratio of Victoria (yellow) to Yamagata (purple) sequences in each segment from the primary dataset over time. Black lines indicate where this ratio lies in the larger secondary dataset. Numbers at the top of the figure show the total number of genomes available for each influenza season in the primary dataset comprised of 452 genomes from which the ratio was calculated, while the numbers in brackets correspond to numbers of sequences in the larger secondary genomes dataset. Numbers at the bottom are influenza seasons from the 1987/1988 (87/88) season to the 2011/2012 season. Yamagata lineage of PA, NP, NA and MP segments and Victoria lineage of the NS segment eventually become fixed (in the population genetics sense of the word) in the influenza B population. PB1, PB2 and HA segments maintain separate Victoria and Yamagata lineages.

We reconstructed reassortment events that were detected by using lineage labels. Figure 4.6 focuses only on inter-lineage reassortments that have occurred after 1990. We identify five major (in terms of persistence) reassortant genome constellations (given in order PB1-PB2-PA-HA-NP-NA-MP-NS with prime (') indicating independently acquired segments) circulating between 1992 and 2011 (Figure 4.6):

- B/Alaska/12/1996-like (Y-Y-Y-Y-Y-Y-Y-V)
- B/Nanchang/2/1997-like (V-V-Y-V-Y-V-Y-V)
- B/Iowa/03/2002-like (V-V-Y'-V-Y-Y-Y'-V')

- B/California/NHRC0001/2006-like (V-V-Y-V-Y'-Y-Y'-V')
- B/Brisbane/33/2008-like (V-V-Y-V-Y'-Y-Y-V)

In a previous study B/Alaska/12/1996-like, B/Nanchang/2/1997-like and B/Iowa/03/2002-like constellations were observed (Chen and Holmes, 2008), but sequences from B/California/NHRC0001/2006-like and B/Brisbane/33/2008-like constellations were not available at the time. In their study Chen and Holmes (2008) also recovered the co-assortment pattern of PB1, PB2 and HA lineages, but did not remark upon it. Of these five constellations four (B/Nanchang/2/1997-like, B/Iowa/03/2002-like, B/California/NHRC0001/2006-like and B/Brisbane/33/2008-like) are derived from introgression of Yamagata lineage segments into Victoria lineage PB1-PB2-HA background, with only one (B/Alaska/12/1996-like) resulting from introgression of Victoria lineage NS segment into an entirely Yamagata derived background. All five inter-lineage reassortment events described here are marked by the preservation of either entirely Victoria or Yamagata derived PB1-PB2-HA segments. Figure 4.6 also shows that reassorting segments appear to evolve with a considerable degree of autonomy. For example, the NP lineage that entered a largely Victoria lineage derived genome and gave rise to the B/Nanchang/2/1997-like isolates continued circulating until 2010, even though the other segments it co-assorted with in 1995 – 1996 (PA and MP) went extinct following the next round of reassortment that led to the rise of B/Iowa/03/2002-like genome constellations. A more extreme example is the NS segment, where a Vic sub-lineage was reassorted into an entirely Yam background (B/Alaska/12/1996-like) in 1994–1995, then reassorted back into a mostly Vic background some 5 years later (B/Iowa/03/2002-like) where it has replaced the 'original' Vic sub-lineage (see Figure 4.6).

We observe that in all five successful inter-lineage reassortment events shown in Figure 4.6, none break up the PB1-PB2-HA complex. This is an unlikely outcome – the probability of not breaking up PB1-PB2-HA across five reassortment events is $p = (\frac{2^5 \times 2 - 2}{2^8 - 2})^5 = 0.0009$, where reassortment events are considered to sample from the Vic and Yam lineages at random for each of the eight segments. If we correct for multiple testing with the assumption that co-assortment of any three segments is of interest we find that the probability of not breaking up an arbitrary set of three segments across five reassortment events is $p = \binom{8}{3} \times (\frac{2^5 \times 2 - 2}{2^8 - 2})^5 = 0.0485$.

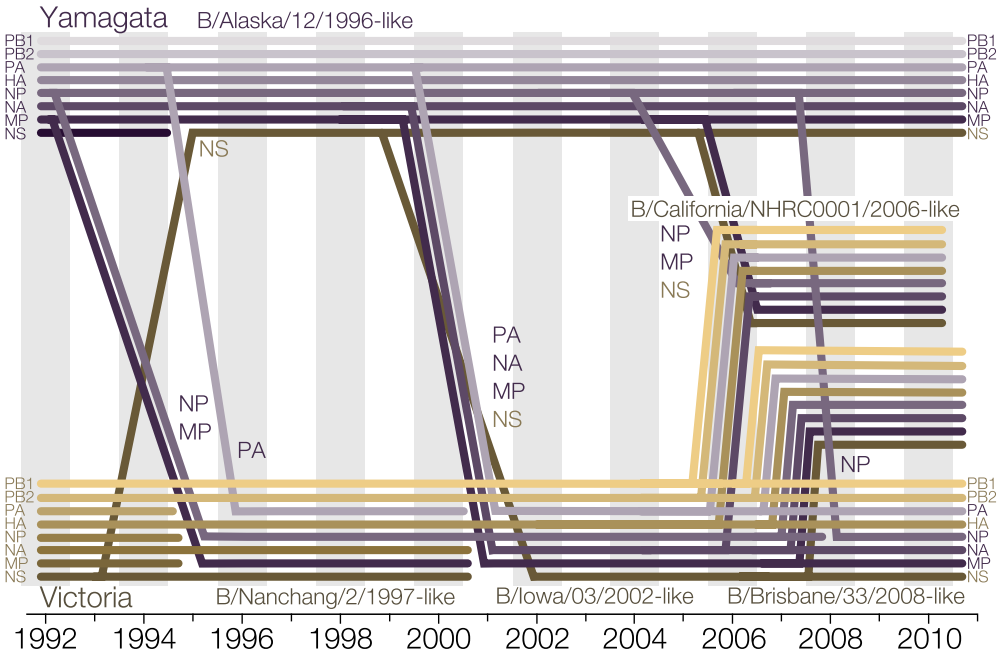


FIGURE 4.6: **Schematic plot of reconstructed reassortments between Victoria and Yamagata lineage segments of influenza B virus.** Lineages that coassort in genomes are represented by 8 parallel lines, with lineages that derive from the original Victoria clade colored yellow/brown and lineages that derive from the original Yamagata clade colored lilac/purple. Inter-lineage reassortment events are indicated by lines entering a different genome. The angle of incoming lineages represents uncertainty in the timing of the event (mean date of the reassortant node and its parent node). Lineage extinction dates are not shown accurately.

Although the vast majority of influenza B isolates possess either Vic or Yam lineage derived PB1-PB2-HA complexes, on rare occasions mixed-lineage PB1-PB2-HA constellations emerge. Figure 4.7 shows the sum of branch lengths which were labelled as having entirely Vic, entirely Yam or mixed-lineage PB1, PB2 and HA segments. Due to lack of reassortment between Vic and Yam lineages of PB1, PB2 and HA (Figure 4.4) since 1997 all segments have spent significantly longer periods of evolutionary time with either entirely Vic-derived or entirely Yam-derived than with mixed-lineage PB1, PB2 and HA constellations (Figure 4.7). We have identified three instances of mixed-lineage PB1-PB2-HA reassortants from the primary dataset with the following PB1-PB2-HA constellations: VVY (B/Bangkok/163/1990-like, 13 sequences isolated 1990 – 5 Jan 1995), VYV (B/Nanchang/630/1994-like, two sequences isolated 1994 – 1996) and VYY (B/New York/24/1993-like, two sequences isolated 8 Jan 1993 –

1994). We detected two new reassortant lineages when investigating the larger secondary dataset – B/Waikato/6/2005-like viruses with PB1-PB2-HA constellation YYV (17 sequences isolated 9 May – 12 October in 2005) and B/Malaysia/1829782/2007 with PB1-PB2-HA constellation YVY (one sequence isolated 2 August 2007).

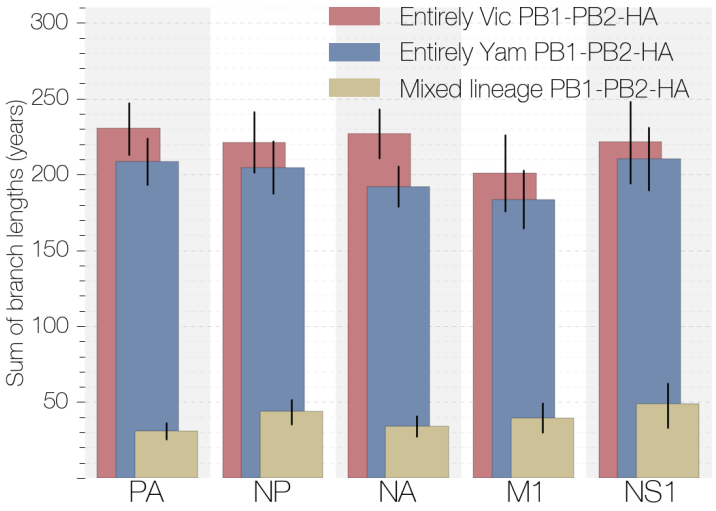


FIGURE 4.7: **Amount of evolutionary time each segment has spent under different PB1-PB2-HA constellations.** All segments have spent significantly more of their history with entirely Vic or entirely Yam-derived PB1-PB2-HA complexes. All vertical lines indicating uncertainty are 95% highest posterior densities (HPDs).

Analysis of reassortment properties

We attempted to quantify the temporal discordance between lineages reassorting into new genomic constellations. If one were able to recover an influenza 'species tree', including admixture/reassortment events, it would be possible to estimate the reassortment or recombination 'distance', which is the time between a split in the species tree in the past and a reassortment event (see Figure B.17). Although we do not find evidence of differences in total number of reassortments between segments (see figure B.5), we find support for a reassortment 'distance' effect, in which a pair of tips on one segment has a different TMRCA from the same pair of tips on a different segment. The summary statistic we use that reflects this difference in TMRCAs, δ_{TMRCA}, is most sensitive when only one of the two trees being compared loses diversity via reassortment and the other acts like a proxy for the 'species tree'. We normalize our δ_{TMRCA} comparisons to arrive at Δ_{TMRCA}, which accounts for uncertainty in tree topology (see Methods). Figure 4.8 shows Δ_{TMRCA} values for all pairs of trees. Most segment pairs show very

low values for this statistic with $\Delta_{\text{TMRCA}} \approx 0.1$, indicating that δ_{TMRCA} measurements between replicate posterior samples from the same segment are up to 10 times smaller than δ_{TMRCA} values between posterior samples from different segments. PB1, PB2 and HA trees, on the other hand, exhibit Δ_{TMRCA} values that are much higher. This shows that TMRCA differences between trees of PB1, PB2 and HA segments are, though noisy, occasionally very similar to uncertainty in tip-to-tip TMRCAs between replicate analyses of these segments.

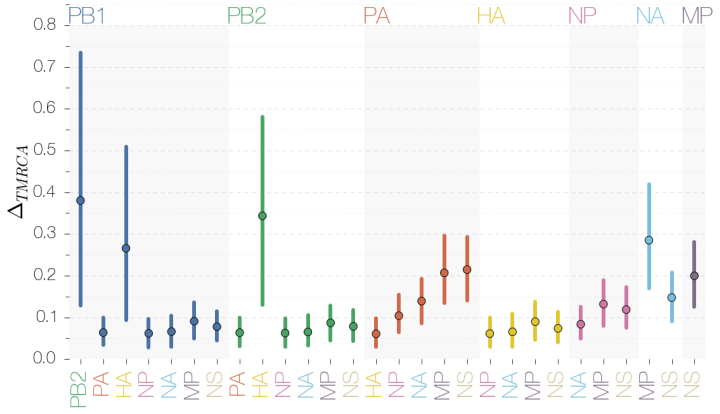


FIGURE 4.8: Δ_{TMRCA} **statistics for different segment pairs.** PB1, PB2 and HA trees exhibit reciprocally highly similar TMRCAs, unlike most other pairwise comparisons. All vertical lines indicating uncertainty are 95% highest posterior densities (HPDs).

Discussion

Linkage between PB1, PB2 and HA gene segments

In this paper we show that the PB1, PB2 and HA segments of influenza B viruses are the only ones that have continuously maintained separate Vic and Yam lineages, while other segments have fixed either Vic or Yam lineages (Figures 4.2, 4.5 and 4.6). Evidence suggests that this is a result of prolonged lack of reassortment between Vic and Yam lineages in PB1, PB2 and HA (Figure 4.4) which possess co-assorting sequences detectable as high linkage disequilibrium (Figure B.1). The vast majority of the sampled evolutionary history of each segment of influenza B viruses since the split of Vic and Yam lineages has been spent in association with either completely Victoria or completely Yamagata lineage derived PB1-PB2-HA complexes (Figure 4.7), suggesting that having 'pure' lineage PB1-PB2-HA complexes is important for whole-genome

fitness. We propose that this pattern of coassortment is due to the action of selection and not simply biased or rare reassortment.

The origin of the strong genetic linkage between PB1, PB2 and HA segments remains unclear. We believe there are two alternative, but similar explanations for the origins of the strong genetic linkage between these segments: mutation-driven co-evolution (Presgraves, 2010) and Dobzhansky-Muller incompatibility (Dobzhansky, 1937; Muller, 1942). Mutation-driven co-evolution (Presgraves, 2010), has been suggested to be the cause of hybrid dysfunction in *Saccharomyces* hybrids (Lee et al., 2008), and evolves as a by-product of adaptation. If one or the other influenza B lineage has undergone adaptation we might expect these changes to be beneficial in its native background and incompatible with a foreign background. Dobzhansky-Muller incompatibility operates in a similar way, but the main difference from the scenario described earlier is that the incompatible alleles are neutral or nearly neutral in their native background and become deleterious or lethal when combined with non-native backgrounds. Emergence of DM incompatibility is aided by geographic isolation. Interestingly, the Victoria lineage of HA was restricted to eastern Asia between 1992 and 2000 (Nerome et al., 1998; Shaw et al., 2002), offering ample time for the budding Victoria lineage to accumulate alleles causing reassortment incompatibility. However, without more genomic data from the past, it is difficult to estimate to what extent influenza B virus population structure contributed to the development of the current segment linkage.

Potential mechanisms for reassortment incompatibility

Unfortunately, the limited amount of genomic data available for the early years of the Vic–Yam split precludes any attempts of answering whether selection or drift have led to the current linkage of PB1, PB2 and HA segments. Although the origins of the linkage between these three segments might be difficult to explain, we can speculate on the nature of reassortment incompatibility. For example, it is intuitive for why this might be the case for PB1 and PB2: both proteins interact directly as part of the RNA-dependent RNA polymerase heterotrimer. Indeed, we observe that PB1–PB2 reassortants are the rarest and least persistent among mixed-lineage PB1-PB2-HA strains and have not been isolated in great numbers. In fact most reassortants breaking

the PB1-PB2-HA complex apart have occured in the past, close to the split of Vic and Yam lineages and have become very rare since.

There is some evidence that the linkage between PB1 and HA might not be a phenomenon restricted to influenza B viruses. It has been established that at least for the 1957 and the 1968 influenza pandemics, caused by A/H2N2 and A/H3N2 subtypes, respectively, the viruses responsible were reassortants possessing PB1 and HA segments derived from avian influenza A viruses (Kawaoka et al., 1989). In addition, outdated techniques for producing vaccine seed strains through selection for HA-NA reassortants often yielded PB1-HA-NA reassortants as a side-effect (Bergeron et al., 2010; Fulvini et al., 2011). Recent experiments have found that the presence or absence of a 'foreign' PB1 segment can have dramatic effects on HA concentration on the surface of virions and total virion production (Cobbin et al., 2013). However, there have been reassortant influenza A viruses circulating for prolonged periods of time in humans that did have disparate PB1 and HA segments, *e.g.* H1N2 outbreaks in 2001 (Gregory et al., 2002) and H1N1/09 in 2009 (Smith et al., 2009).

We believe that the association between PB1, PB2 and HA segments should be relatively straightforward to explore in the lab. Reverse genetics systems have been developed for influenza B viruses (Hoffmann et al., 2002), which would allow the creation of artificial reassortants. Based on the frequency and persistence times of different reassortant classes we have observed, we expect a hierarchy of reassortant fitness starting with PB1+PB2+HA reassortants which should be the most fit, followed by PB1+2/HA, then PB1+HA/PB2 and finally PB2+HA/PB1 reassortants with the lowest fitness. We believe that this is the most direct approach to unravelling the mechanism responsible for the linkage within the PB1-PB2-HA complex.

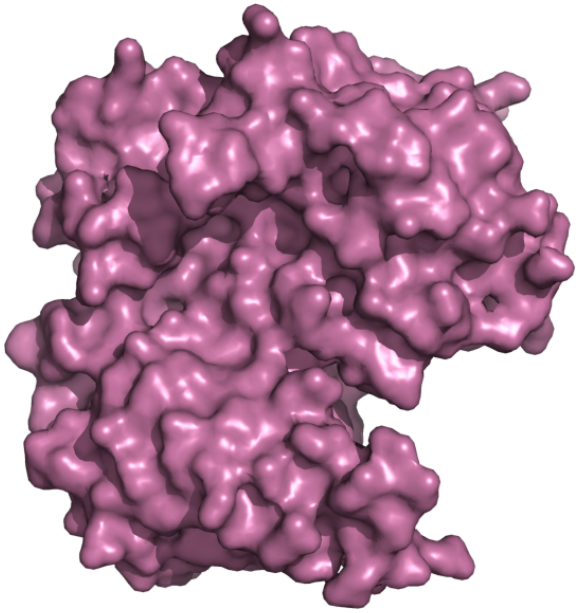
Will influenza B viruses speciate?

We suggest that the preservation of two PB1-PB2-HA complex lineages is similar to genomic speciation islands, where small numbers of genes resist being homogenized through gene flow (Turner et al., 2005). In this context, we see three potential paths of evolution for influenza B viruses. If more segments get recruited to the PB1-PB2-HA complex, the process could continue until 'speciation' occurs in which none of the

segments are able to reassort across the Victoria–Yamagata lineage boundary. Alternatively, the influenza B genome could continue to be homogenized via gene flow with the exception of PB1, PB2 and HA segments or one of the two PB1-PB2-HA complexes could go extinct, marking the return of single-strain dynamics in the influenza B virus population. The eventual fate of influenza B viruses will likely be determined by the combined effects of reassortment frequency and the strength of epistatic interactions between segments.

CHAPTER 5

REASSORTMENT BETWEEN CLOSELY RELATED INFLUENZA B LINEAGES AND THEIR DISPERSAL IN SCOTLAND



Anticlimax is, of course, the warp and way of things. Real life seldom structures a decent denouement.

– Dan Simmons, *Hyperion*

Data collection was performed by Donald B. Smith, Karina Salvatierra, Rory Gunson and Heli Harvala. The analysis and interpretation of data were done by myself.

Introduction

In a previous chapter (chapter 4) I have shown that genomes of currently circulating influenza B viruses, which are thought to be endemic human pathogens, contain co-segregating segment lineages. The genome-wide split of influenza B virus segments into two co-circulating lineages designated as Victoria (or Vic) and Yamagata (or Yam) lineages occured some time in the 1980s on the basis of haemagglutination inhibition assays (Rota et al., 1990), followed by often complicated patterns of reassortment between the two genomic lineages (Lindstrom et al., 1999; Chen and Holmes, 2008). However, over the decades since the origin of the two lineages most of the influenza B virus genome had become homogenised, as segments from one lineage introgressed into the genomic background of the other, resulting in reduction of genetic diversity of influenza B viruses, with the exception of PB1, PB2 and HA segments (Dudas et al., 2015). These three segments of the influenza B virus genome do not appear to reassort readily with respect to each other whilst circulating in humans. As a result, modern influenza B virus genomes or reassortants thereof almost always contain PB1, PB2 and HA derived entirely from either Victoria or Yamagata lineage.

In the previous chapter I have hypothesised that there are two possible mechanisms that could explain the linkage between PB1, PB2 and HA segments. These explanations would be equivalent to two broad classes of reproductive isolation mechanisms. If speciation literature language were to be applied it would be equivalent to pre-zygotic and post-zygotic isolation mechanisms. If influenza B virus segments contain packaging signals like those described in influenza A viruses (Gog et al., 2007), especially if packaging signals are segment-specific, then it would be reasonable to suspect that the three co-segregating segments simply cannot reassort *in vivo*, despite opportunities to do so. It is equivalent to a pre-zygotic isolation mechanism from speciation literature, where different lineages are prevented from exchanging genetic material before the production of hybrid progeny. If influenza B virus segments are prevented from reassorting during

co-infection as a result of incompatibilities in segment packaging, then this should be observable as a reduced reassortment rate between PB1, PB2 and HA segments compared to the rest of the genome. In chapter 4 I have shown, by using subtree prune and regraft (SPR) distances between phylogenetic trees of the influenza B virus segments, that there are no apparent reductions in reassortment rate between any pair of influenza B virus segments. However, it was also apparent that approximate SPR distances are a noisy measure, partly because an approximation was used and partly because phylogenies sampled over the course of MCMC are quite variable in topology.

The alternative hypothesis for the linkage between Vic and Yam PB1-PB2-HA complexes is that influenza B viruses bearing mixed-lineage PB1-PB2-HA complexes have inferior fitness, through any number of genetic effects, such as mutation-driven co-evolution (Presgraves, 2010) or Dobzhansky-Muller incompatibilities (Dobzhansky, 1937; Muller, 1942). This would be equivalent to a post-zygotic isolation mechanism, where mixed-lineage PB1-PB2-HA reassortants are readily produced upon co-infection of a host cell but fail at a later stage. Based on the isolation of a number of influenza B virus genomes with mixed-lineage PB1-PB2-HA segments over the years and the short duration of their circulation (Dudas et al., 2015) a reassortment incompatibility mechanism relying on selection against readily produced reassortant progeny seems a more likely explanation than inherent biases in genome packaging.

However, additional analyses could be carried out to further differentiate between the two hypotheses of PB1-PB2-HA linkage. If PB1, PB2 and HA are constrained in the extent to which they reassort and in the absence of lineage-specific packaging signals or any other lineage effect, then biased reassortment between PB1, PB2 and HA segments should be detectable within each lineage. To this end we analyse two sets of influenza B virus sequence data and estimate the numbers of reassortments between HA and NA segments that have taken place within Victoria and Yamagata lineages, as well as HA, NA, PB1 and PB2 segments. I include HA and NA sequences sampled densely between 2007/2008 and 2012/2013 influenza seasons in Scotland (see figure 5.1) for the first part of the analysis. If reassortment is sufficiently frequent and detectable, then it would be possible to determine whether reassortment occurs locally or whether reassortant viruses originate elsewhere and spread globally. The second dataset comprised of HA, NA, PB1 and PB2 sequences contains the subset of HA and NA sequences for which PB1

and PB2 segments are available. This was done to avoid potential biases in country-specific sequencing capacity, since many laboratories globally usually sequence HA and NA segments, but full genome sequencing, which is often how PB1 and PB2 sequences are recovered, is performed in much smaller numbers and in countries with excellent sequencing capacity.

Since detailed geographic information was available for HA sequences from Scottish isolates a secondary analysis was performed to investigate the geographic movements of influenza B viruses in Scotland. Influenza B viruses have been shown to exhibit geographic structure much stronger than that of influenza A viruses, as a result of reduced mobility of children, who represent the most commonly infected age group (Bedford et al., 2015). This might present a unique opportunity for reassortment, since geographic confinement could yield distinctive local reassortants. Additionally, analyses in the past have suggested that locations further from the equator tend act as human influenza sinks, with viruses circulating near the equator seeding epidemics in the northern and southern hemispheres during their respective winters, which die out, only to be reseeded again once the conditions are conducive to their transmission (Nelson et al., 2007; Russell et al., 2008; Bedford et al., 2010). Investigating influenza B virus circulation in Scotland in detail, in the context of global influenza B virus diversity, allows to test the claim that the two hemispheres act as influenza virus sinks, as well as investigating the migration of a human virus within an area where air travel is not the primary mode of transport.

Methods

Scottish influenza B virus sequence data

134 influenza B viruses were collected by the National Healthcare Service (NHS) boards in Scotland. All 134 viruses had their HA segment sequenced as well as the NA segment of a subset of 31 viruses. For more information about the sequencing protocol see appendix C. Some of the sequences were of poor quality, with some sequences having excessive numbers of ambiguous base calls or potential sequencing errors and were not submitted to GenBank as a result. This casts some doubts on the analyses presented

here, although sequencing errors should be unique to each sequence and thus should only affect external branches by making them longer in substitution trees or leading to inference of a higher evolutionary rate on those branches in a relaxed molecular clock analysis for temporally calibrated phylogenies. However, it is important to note that poor quality was only an issue for a relatively small subset of sequences, mostly from the Lothian NHS board.

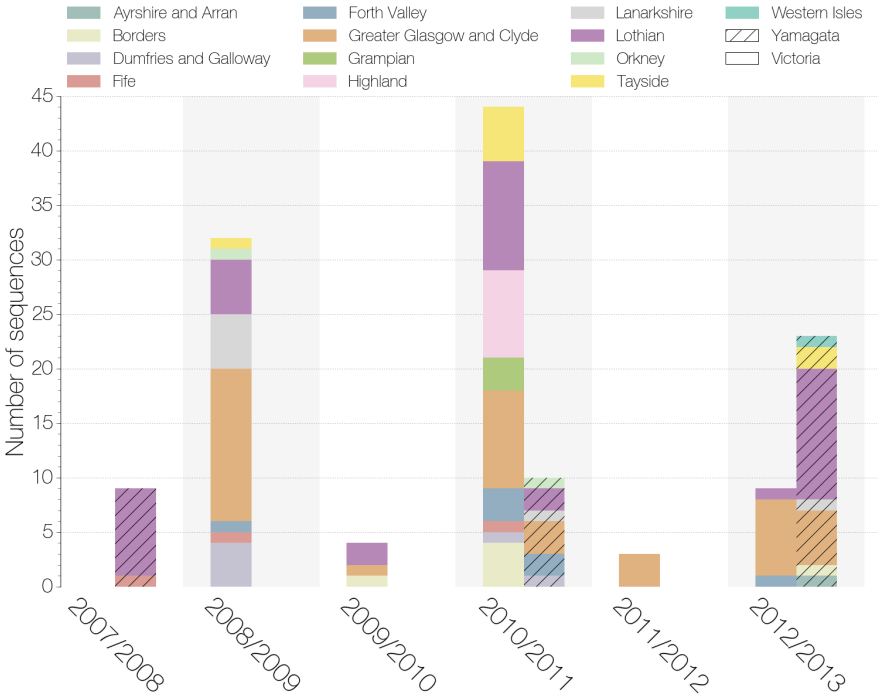


FIGURE 5.1: **Numbers of isolates from Scotland sequenced for this study.** The numbers of isolates from Scotland for which sequence data are available split by influenza season, lineage and location. Locations are coloured by National Healthcare Service board where viruses were isolated and sequenced. Hatched bars indicate Yamagata lineage sequences, Victoria lineage is displayed in solid colours.

Background datasets

Three datasets were collated from GenBank to provide a context for sequences from this study. The first dataset was composed of strains for which both the HA and the NA sequences were available and were used to infer HA-NA reassortants. We split these sequences into Victoria and Yamagata lineages, based on the lineage of HA. Additionally for Victoria lineage sequences we only included sequences in the dataset that were descendants of the major NA inter-lineage reassortment event that took place around 2001 (Dudas et al., 2015), giving rise to B/Iowa/03/2002-like viruses, which are

Victoria lineage HA-bearing influenza B viruses with a Yamagata lineage NA. This was done by excluding isolates with Victoria lineage NA segments. After the sequences were split into Victoria and Yamagata categories we sampled randomly from both to give the final datasets. We chose a sampling frequency of six sequences per month for Victoria lineage and 10 sequences per month for Yamagata lineage sequences to give similar-sized datasets of 499 and 474 HA-NA pairs, respectively (for sampling density see figure C.1). This was repeated one more time to give us two replicate datasets for both Victoria and Yamagata lineages, albeit with a considerable overlap between the replicates: 386 and 443 shared sequences between replicates for Victoria and Yamagata lineages, respectively. This dataset was used to find reassortments between HA and NA segments, the most important segments responsible for antigenicity of influenza B viruses.

The second dataset was comprised of sequences derived from the first dataset for which PB1 and PB2 sequences were also available, to further investigate the reassortment patterns between PB1, PB2, HA and NA segments. We chose to go from a large HA-NA dataset to a smaller PB1-PB2-HA-NA dataset, rather than from a PB1-PB2-HA-NA dataset to an HA-NA dataset in order to avoid sampling biases and to capture as much of the circulating HA and NA genetic diversity as possible. The most worrying sampling bias for this particular dataset could arise as a result of particular countries having the capacity to sequence full influenza B virus genomes. This could bias the entire dataset towards sequences from countries capable of sequencing full influenza B virus genomes, without properly reflecting the global diversity of PB1, PB2, HA and NA segments. Between the two replicate datasets of the Yamagata lineage, composed of strains for which both HA and NA segment sequences were available, replicate dataset number 2 contained more strains for which PB1 and PB2 sequences were also available. For Victoria lineage it was replicate dataset number 1. The final dataset sizes were: 184 PB1, PB2, HA and NA sequences for Yamagata lineage and 234 PB1, PB2, HA and NA sequences for Victoria lineage (for sampling density in this dataset see figure C.2). This dataset was used to test whether the reassortment patterns between HA, PB1 and PB2 described in our previous work (Dudas et al., 2015) also extended to within-lineage dynamics.

The third dataset was comprised entirely of Victoria and Yamagata lineage HA sequences and in contrast to datasets one and two was used to analyse the phylogeography of these viruses within Scotland and in the context of the rest of the world. The sequences for this dataset were collated from GenBank and only included sequences isolated in 2006 or later. We sampled from datasets of each lineage at a frequency of five sequences/month and proportionally to the population size of the country from which they were isolated (for sampling density see figure C.3).

Phylogenetic analyses

The first dataset, comprised of HA-NA pairs of Victoria and Yamagata sequences (with two replicates each), was analyzed in BEAST (Drummond et al., 2012) under the following model: HKY nucleotide substitution model (Hasegawa et al., 1985) with rate heterogeneity (four gamma rate categories) (Yang, 1994), partitioned into codon positions 1+2 and 3, non-parametric Bayesian skyride as the tree prior (Minin et al., 2008), uncorrelated relaxed lognormal molecular clock with tip date calibration (Drummond et al., 2006), with some tip dates, where not precisely known, estimated as a latent variable over the course of the MCMC chain. The choice of demographic tree prior was determined by the size of the datasets. The multilocus Bayesian skygrid (Gill et al., 2013) tree prior would have been preferable for all analyses described herein, since it could improve estimates of common ancestor times in both trees via a shared demographic history, but the sizes of each dataset would prevent efficient MCMC sampling of both phylogenies, which are relatively big (499 and 474 tips for Victoria and Yamagata lineage, respectively). MCMC chains were set up to run for 200 million states and sampled every 20000 steps with each analysis being done in triplicate, giving a total of 24 analyses (two replicates $\times$ Vic/Yam $\times$ HA/NA $\times$ three independent Markov chains).

The second dataset, which included sequences of PB1, PB2, HA and NA segments, was analysed under the same set of parameters as the first analysis, except that the MCMC chain was shortened to 100 million states and sampled every 10000 states. This resulted in a total of 16 analyses (four segments $\times$ Vic/Yam $\times$ two independent Markov chains).

Reassortment identification

Since no widely accepted and biologically relevant lineage divisions exist within Victoria and Yamagata genomes, and because the goal of this chapter is to determine the numbers, rather than properties, of reassortment within influenza B virus lineages I used GiRaF (Nagarajan and Kingsford, 2011) rather than lineage labelling methods, as was done in chapter 4. GiRaF uses splits extracted from posterior distributions of trees generated by Bayesian methods and searches for highly supported but incongruent splits (posterior probability that splits exist and are incongruent >0.5) between two posterior distributions of trees. For the first dataset GiRaF was used to find reassortant clades between HA and NA trees in Victoria and Yamagata categories and in each replicate set of sequences. Cross-checking clades that were shared between replicates allowed us to exclude likely false positives. We consider clades to be false positives if they are identified as reassortant in one replicate dataset, but not the other, despite sharing substantial numbers of sequences.

In addition GiRaF was used to analyse PB1, PB2, HA and NA segments jointly. This approach produces a 'catalog' file which provides information on which segments are reassorting with respect to other segments. For example, "PB1 PB2 HA | NA" would designate a reassortment event where PB1, PB2 and HA trees have a consistent topology, but NA does not and has likely reassorted with resepct to the other three. Due to excessive numbers of possible analysis combinations only one replicate was analysed per Victoria/Yamagata lineage.

Phylogeographic analyses

The third dataset was comprised of HA sequences was run under the same parameters as the first analysis but with the addition of ancestral reconstruction of discrete states using an asymmetric transition matrix (Lemey et al., 2009b) with Bayesian stochastic search variable selection (BSSVS) to estimate significant rates of state transitions. Locations of each sequence were used as discrete states, based on the NHS region of isolation and one additional category, "other" to represent background sequences from outside of Scotland. Scottish sequences came from 13 regional Boards of the National Health Service (NHS) in Scotland: Borders (BORD), Dumfries and Galloway (DG), Fife

(FIFE), Forth Valley (FV), Greater Glasgow and Clyde (GGC), Grampian (GRAM), Highland (HIG), Lothian (LOT), Orkney (ORK), Tayside (TAY), Western Isles (WI) and Ayrshire and Arran (AA). In order to assess the gene flow into and from each location or collection of locations (Scotland) we used 'Markov jump' counts (Minin and Suchard, 2008), which is a computationally tractable method of sampling realizations of a continuous time Markov chain process.

Counting influenza B virus introductions into Scotland

We used trees derived from the third dataset, comprised entirely of HA sequences, to estimate the rate at which influenza B viruses are introduced into Scotland and assume that this is representative of the behaviour of the whole genome. One could imagine that HA segments could be introduced into Scotland and then reassorted into a locally circulating influenza B virus genome, although there is no described evidence of this. We estimated the number of influenza B virus introductions as the number of non-Scottish to Scottish state transitions along branches of the phylogenetic tree. Unlike reassortment numbers, which were inferred *post hoc* from the whole posterior distribution by GiRaF, our estimates of the number of state transitions were integrated over uncertainty in tree topology and other parameters. Having samples from all over Scotland also allowed us to investigate, to a limited extent, the movement of the virus within Scotland. We employ Markov jumps (Minin and Suchard, 2008), a computationally tractable way of counting labeled transitions, which in our case are switches from one NHS board/location to another, as well as transitions between locations in Scotland versus the rest of the world.

Results

Victoria lineage viruses reassort more frequently than Yamagata lineage viruses

Using GiRaF we discovered 8-16 reassortments (depending on replicate) within Victoria and 5-6 reassortments within Yamagata lineage in the primary dataset. Because we used two replicate datasets for this analysis we were able to cross-reference GiRaF-identified

reassortant clades and keep only those that were identified in both replicates. After cross-referencing we were left with seven clades of Victoria lineage and four clades of Yamagata lineage HA-NA reassortants (see figure 5.2). Overall, this suggests potential differences in reassortment frequency and reassortant success within Victoria versus within Yamagata lineage.

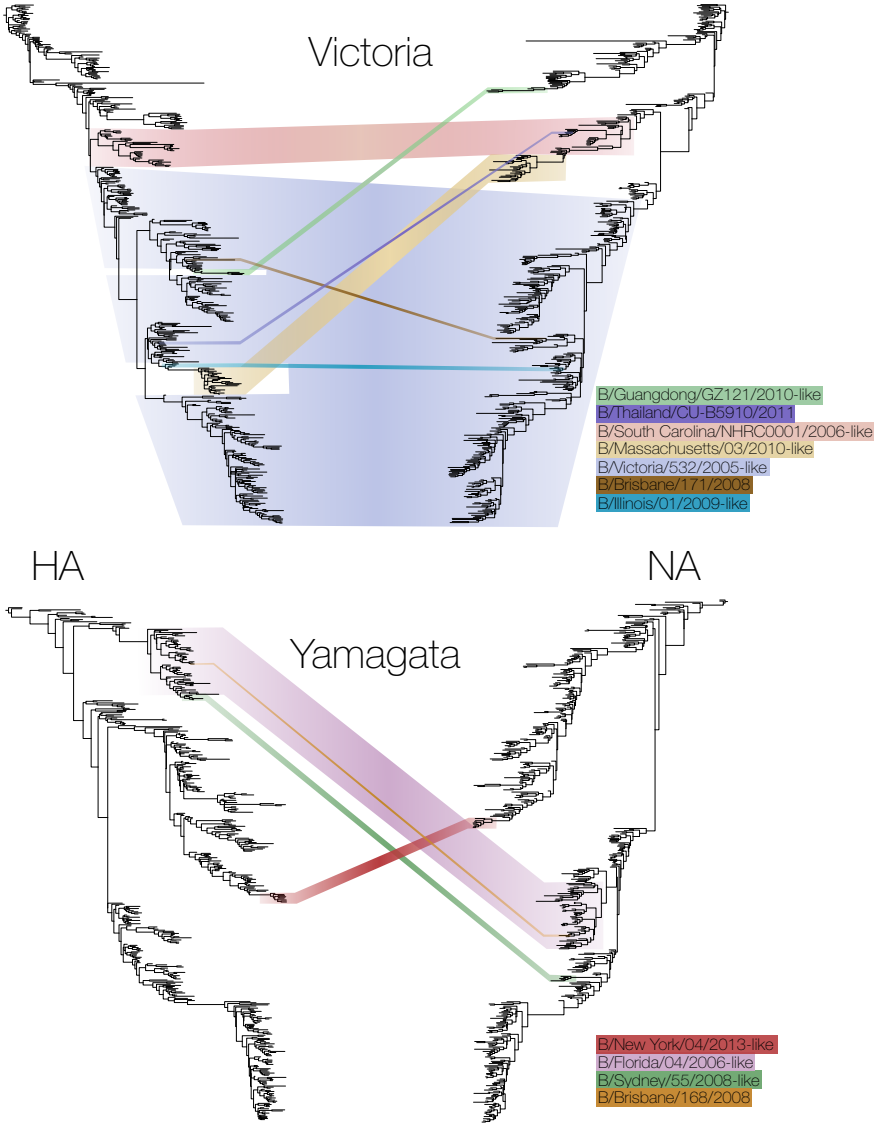


FIGURE 5.2: **Inferred HA-NA reassortant clades.** GiRaF-identified and confirmed reassortant clades and their phylogenetic positions in HA (left) and NA (right) trees of Victoria (top) and Yamagata (bottom) lineages. Victoria lineage was identified to have undergone seven within-lineage HA-NA reassortments whereas Yamagata lineage was inferred to have undergone four such reassortments. Reassortments within Victoria lineage also appear to be more successful on average, with reassortant clades possessing larger numbers of isolates.

We also observed a higher number of confirmed reassortants within Victoria lineage in

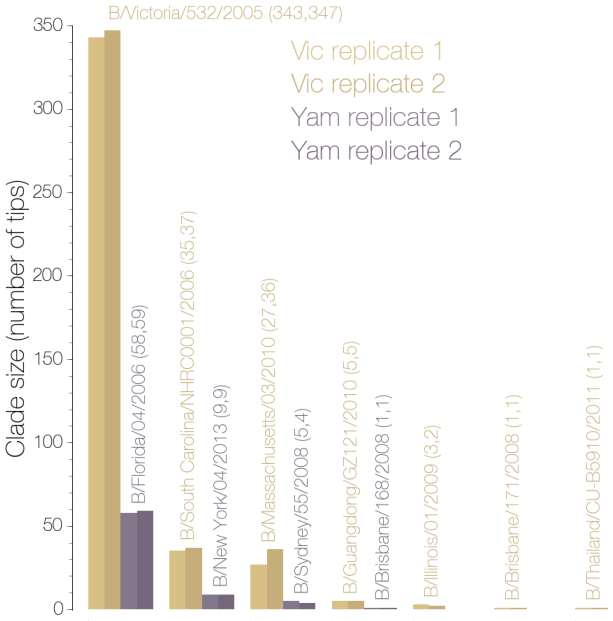


FIGURE 5.3: **Summary of HA-NA reassortant clades.** The figure shows the clade sizes of HA-NA reassortants, as identified by GiRaF, as well as the names of the earliest isolates within each reassortant clade. Numbers in brackets following clade designations are clade sizes of each reassortant in replicates 1 and 2, respectively. We verified seven HA-NA reassortments within Victoria and four within Yamagata lineage by cross-referencing GiRaF-identified clades between replicate datasets.

the secondary (PB1, PB2, HA and NA) dataset. There was a total of eight reported PB1-PB2-HA-NA reassortants within Victoria lineage versus five within Yamagata lineage. Unfortunately, since few isolates with PB1 and PB2 segment sequences are available it is difficult to verify these clades as reassortants using a replicate dataset. We do note, however, that the HA segment appears to be reassorted more frequently on its own (5/13 events) or in association with NA (5/13 events), followed by the NA segment on its own (2/13 events) (table 5.1). Only one case of a non-HA and non-NA segment being reassorted was identified - the single Yamagata lineage isolate B/Victoria/516/2004 with a reassorted PB2 segment.

Seeding and diffusion of influenza B epidemics in Scotland

Phylogeographic analyses suggest that influenza B viruses can sustain transmission within Scotland for long enough to detect diffusion of the virus. Sequences isolated by the Greater Glasgow and Clyde (GGC) and Lothian (LOT) NHS boards appear to be especially important for initial introductions of influenza B viruses into Scotland, given

lineage	reassortant class		type isolate	clade size
Vic	PB1, PB2, HA‡	NA	B/South Carolina/NHRC0001/2006	27
Vic	PB1, PB2, NA	HA	B/Victoria/210/2007[2]	11
Vic	PB1, PB2	HA, NA‡	B/Victoria/508/2008†	6
Vic	PB1, PB2	HA‡, NA	B/Victoria/508/2008†	2
Vic	PB1, PB2	HA‡, NA	B/Taiwan/70690/2006	2
Vic	PB1, PB2	HA‡, NA	B/Victoria/508/2008	1
Vic	PB1, PB2, HA	NA	B/Mississippi/UR-06-0477/2007[1]	1
Vic	PB1, PB2, NA	HA	B/Brisbane/171/2008	1
Yam	PB1, PB2, NA	HA	B/Auckland/1/2007[2]	34
Yam	PB1, PB2, NA	HA	B/Sydney/55/2008	3
Yam	PB2, NA	HA	B/Sydney/102/2008[2]	2
Yam	PB1, HA, NA‡	PB2	B/Waikato/70/2005	1
Yam	PB1, PB2	HA, NA	B/Victoria/516/2004	1

TABLE 5.1: **Summary of verified PB1-PB2-HA-NA reassortant clades.** The table shows the sizes of clades in the secondary dataset that were identified as PB1-PB2-HA-NA reassortants by GiRaF, as well as the names of the earliest isolates within each reassortant clade. Candidate clades identified by GiRaF were confirmed by visual inspection of the phylogenetic trees. ‡Segment not identified by GiRaF as reassortant, but inferred from the primary HA-NA dataset.
†Clade is nested within another reassortant clade and is designated by the same earliest isolate. [1]B/South Carolina/NHRC0001/2006-like in primary HA-NA dataset. [2]B/Florida/04/2006-like in primary HA-NA dataset.

that both have the highest numbers of inferred introductions from the rest of the world (figures 5.6 and 5.9). Greater Glasgow and Clyde, and Lothian NHS boards service the two largest cities in Scotland (Glasgow and Edinburgh, respectively). Both cities have international airports and so their roles as recipients of influenza B virus lineages are not surprising. There are also indications that both GGC and LOT areas are important for subsequent spread of viruses to the rest of Scotland (figures 5.4 and 5.7). Somewhat surprisingly influenza B virus lineages sampled in GGC and LOT usually comprise distinct but simultaneous epidemics.

There are subtle differences between the two lineages of influenza B viruses in their geographic diffusion. Victoria lineage specifically appears to be predisposed to geographic spread with more significant rates being inferred compared to Yamagata lineage (tables 5.2 and 5.3), although more sequences are also available for Victoria, rather than Yamagata lineage (figure 5.1). In addition, Victoria lineage viruses appear to be introduced into Scotland at a relatively higher rate compared to Yamagata lineage viruses (tables 5.2 and 5.3).

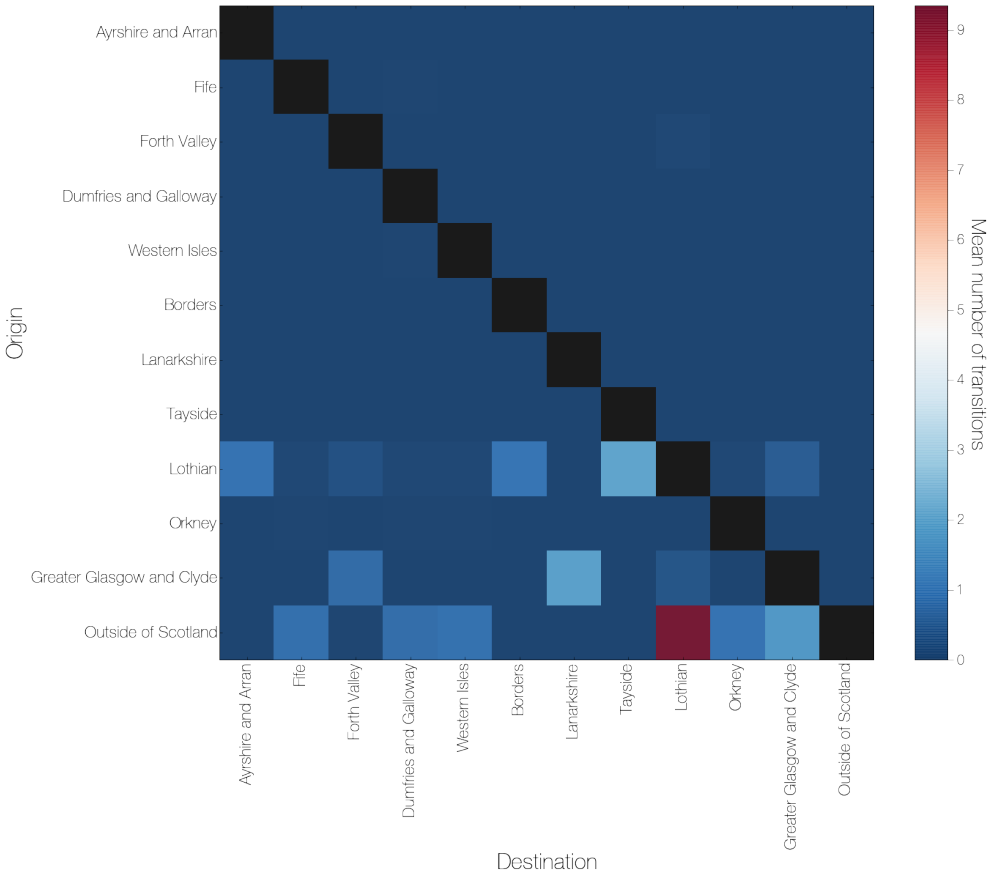


FIGURE 5.4: **Inferred movements of Yamagata lineage based on HA sequences.** Heatmap shows the mean number of inferred transitions between NHS Boards in Scotland. Data shown is from replicate 1.

We also observe what appears to be considerable periods of persistence of influenza B virus lineages within Scotland. Many lineages of both Victoria and Yamagata lineage HAs cluster together into clades that share a common ancestor at least 1 year ago (figures 5.5 and 5.8). This is especially true for Yamagata lineage HA sequences, which have relatively few introductions into Scotland and many more clades that have been present in the area of the Lothian NHS board for prolonged periods of time (figure 5.5). Victoria lineage, on the other hand, exhibits reduced persistence within locations (figure 5.8), but higher migration rates.

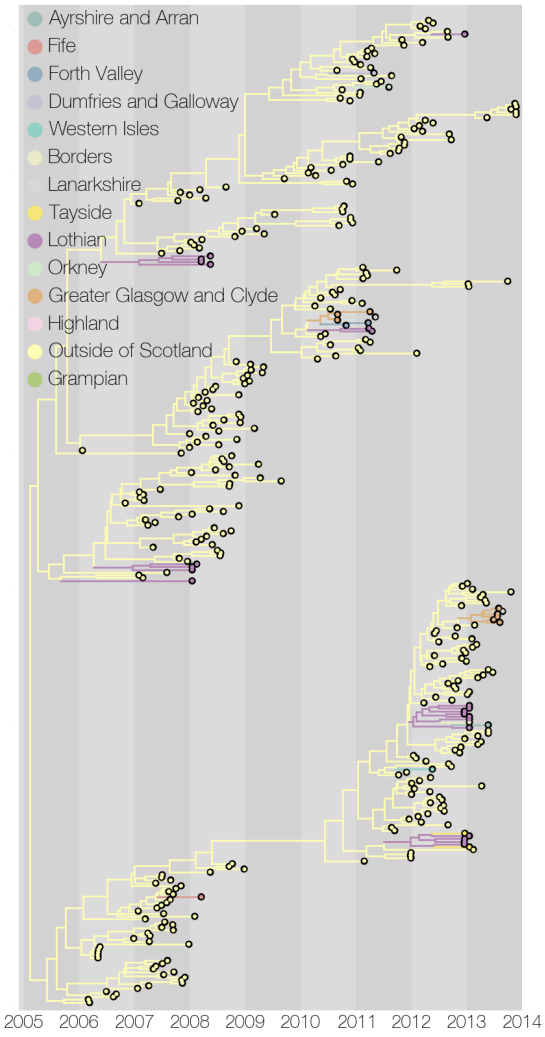


FIGURE 5.5: **Temporally calibrated phylogeny of Yamagata lineage HA coloured by inferred location.** Most of Yamagata lineage evolution appears to take place outside of Scotland. However, there are numerous lineages that appear to be persistent, to some degree, within the Lothian area. In addition, there is evidence that there are two co-circulating and divergent lineages of Yamagata lineage HA segments. Tree shown is from replicate 1.

Discussion

Within-lineage reassortment dynamics of influenza B viruses

The success of detecting recombination or reassortment largely depends on two parameters: genetic diversity and the rate at which recombination or reassortment occur. Best results are achieved under high genetic diversity and low recombination/reassortment rates (Posada and Crandall, 2001). In this paper we have tried to identify reassortant

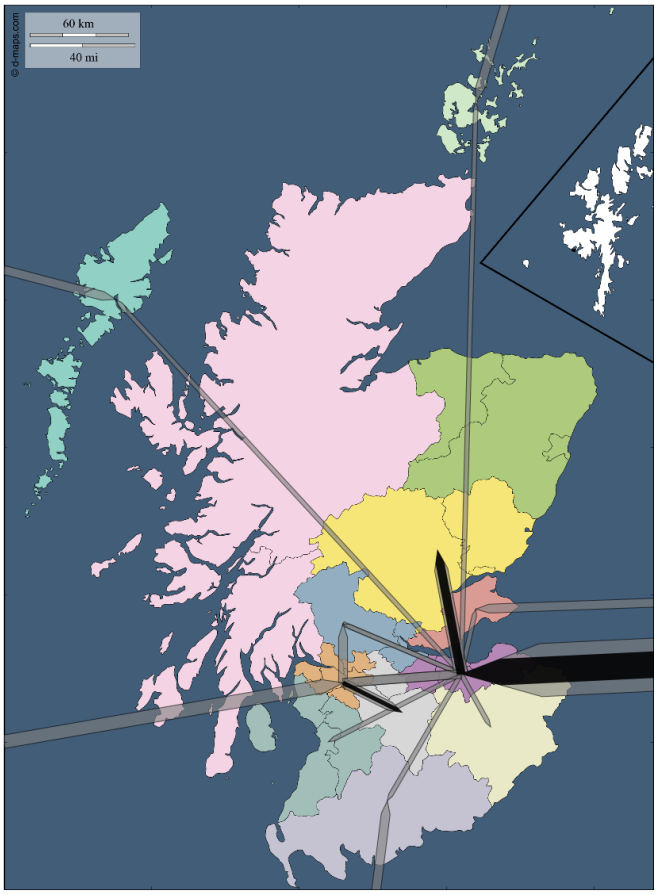


FIGURE 5.6: **Map showing inferred Yamagata lineage movements based on HA.** Arrows indicate Yamagata lineage HA movements within Scotland. The grey transparent arrow indicates the upper 95% highest posterior density value for the number of virus movements along the path. The black arrow within the grey transparent arrow shows the lower 95% highest poterior density value, if it is absent it indicates that the lower 95% HPD interval includes zero. Arrows originating from outside of Scotland indicate inferred influenza B virus movements from the rest of the world into Scotland. The highest confidence exsists for Yamagata lineage movements from the rest of the world into the Lothian NHS board area, from Lothian to Tayside and from Greater Glasgow and Clyde to Lanarkshire NHS board areas. Map shown is from replicate 1.

clades of influenza B viruses under sub-optimal conditions by focusing on each of the two main influenza B lineages individually.

Despite relatively low levels of genetic diversity present within each influenza B virus lineage we have nonetheless identified and confirmed, by cross-validation of replicate datasets, 11 clades of reassortants where the HA and the NA segment are of different origins. Seven of these clades were detected within Victoria lineage and only four within Yamagata lineage. If we assume that every time there is a reassortment event it occurs in Victoria or Yamagata lineage with equal probability then the binomial probability

Origin	Destination	Events/year	BF
Lothian	Ayrshire and Arran	1.22	>30
Lothian	Forth Valley	1.17	4.54
Lothian	Tayside	1.84	>30
Lothian	Greater Glasgow and Clyde	1.54	4.71
Lothian	Borders	1.24	>30
Greater Glasgow and Clyde	Lanarkshire	2.15	>30
Greater Glasgow and Clyde	Forth Valley	1.50	26.56
Greater Glasgow and Clyde	Lothian	1.49	4.74
Outside of Scotland	Fife	0.47	>30
Outside of Scotland	Dumfries and Galloway	0.47	>30
Outside of Scotland	Western Isles	0.48	>30
Outside of Scotland	Lothian	2.22	>30
Outside of Scotland	Orkney	0.46	>30
Outside of Scotland	Greater Glasgow and Clyde	0.70	>30

TABLE 5.2: **Summary of statistically significant (Bayes factor >3) NHS Board transitions for Yamagata lineage.** Although there are relatively few connections between locations in Scotland where the lower 95% HPD of the number of Yamagata lineage migrations excludes 0, many rates were in fact found to be significant by BSSVS. However, the inferred mean rate (events/year) is still quite low for most transitions. The highest inferred rate connects Greater Glasgow and Clyde with Lanarkshire and areas outside of Scotland with Lothian and Lothian with Tayside, corresponding with the highest numbers of inferred migrations between them (figure 5.6). The cut off for statistical significance was considered to be Bayes factor (BF) above 3. Results are shown from replicate 1.

of observing less than or exactly 4 reassortment events out of 11 is 0.274, which is not statistically significant nor noteworthy. Although not statistically significant, this result joins a number of other studies (Vijaykrishna et al., 2015; Dudas et al., 2015) that have found a disproportionate number of reassortments involving Victoria, rather than Yamagata, lineage. Our previous study (chapter 4) found that inter-lineage reassortment events over long timescales tend to involve Victoria lineage PB1-PB2-HA segments acquiring Yamagata lineage segments (Dudas et al., 2015).

We discovered 13 PB1-PB2-HA-NA reassortant clades in our smaller secondary dataset. Five of these reassortants are known from single isolates, which either reflects their rarity or the small size of the dataset. One striking finding from this analysis was that most reassortment events involve the HA segment (10 out of 13 events), either on its own or in association with the NA segment. There were 2/13 events involving the NA segment and 1/13 event involving the PB2 segment. Due to small numbers of reassortments we are unable to dismiss either the 'selectionist' or the 'reassortment bias' explanations for the linkage beween PB1, PB2 and HA segments with much confidence.

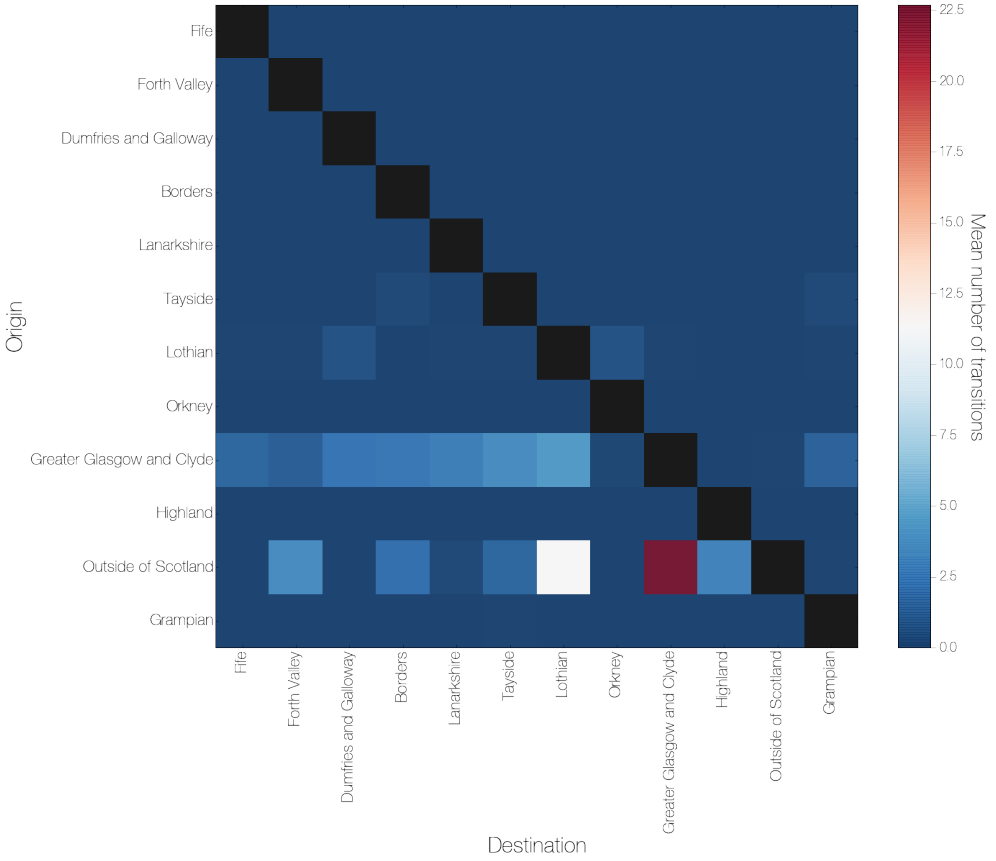


FIGURE 5.7: **Inferred movements of Victoria lineage based on HA sequences.** Heatmap shows the mean number of inferred transitions between NHS Boards in Scotland. Data shown is from replicate 1.

However, the rarity of reassortants separating PB1 and PB2 segments is unexpected and could indicate a lack of power when these two segments are analysed, due to their lower substitution rates. Higher substitution rates increase the resolution of phylogenetic trees by increasing the confidence in each branch and also by improving the ability to distinguish between closely related sequences, which in turn makes reassortments easier to detect.

Contrary to the apparent lack of PB1 and PB2 reassortants, we are seeing what appears to be an excess of HA reassortants. The most straightforward explanation would be that this represents selection for HA reassortants, whereby acquiring antigenically novel HA segments is advantageous to other segments in the genome. This has not been investigated in great detail for any influenza virus, although an example of this has been described by Holmes et al. (2005) where a severe epidemic of influenza A virus subtype H3N2 in 2003–2004 was caused by A/Fujian/411/2002-like viruses that had

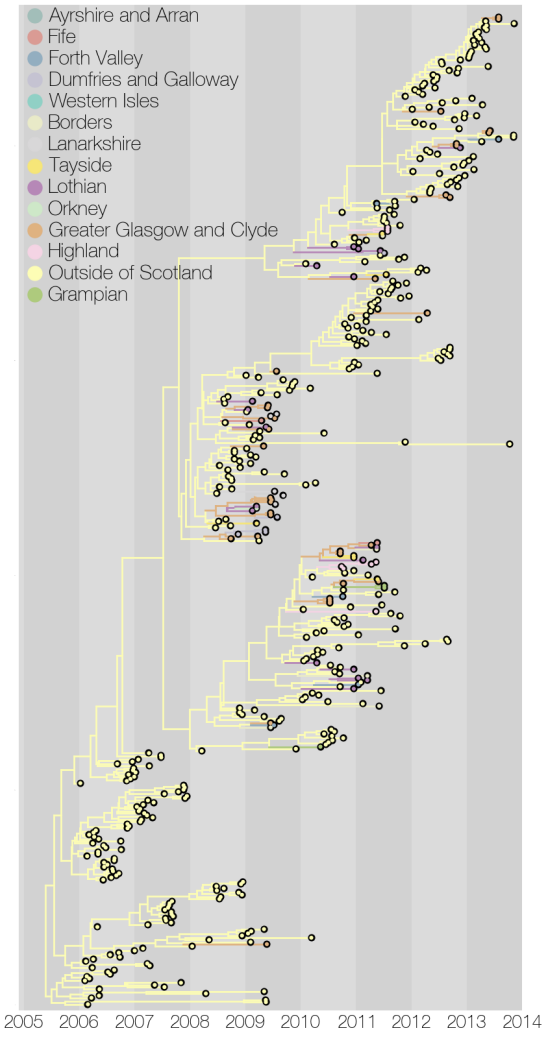


FIGURE 5.8: **Temporally calibrated phylogeny of Victoria lineage HA coloured by inferred location.** Like for Yamagata lineage, most of Victoria lineage evolution appears to take place outside of Scotland. n contrast to Yamagata lineage there are fewer clades that are persistent over long periods of time within Scotland and the overall tree shape suggests more rapid lineage turnover in Victoria lineage compared to Yamagata lineage. Tree shown is from replicate 1.

acquired an HA segment from a minor co-circulating H3N2 lineage.

Local phylogeography of influenza B viruses

Our analyses for both Victoria and Yamagata lineages in Scotland indicate that the rest of the world seems to be the most important source of influenza B viruses, rather than endemic lineages. This agrees with conclusions reached by many other studies looking at the global circulation of influenza viruses (Nelson et al., 2007; Russell et al., 2008; Bedford et al., 2010). Namely, these studies have found that countries at higher

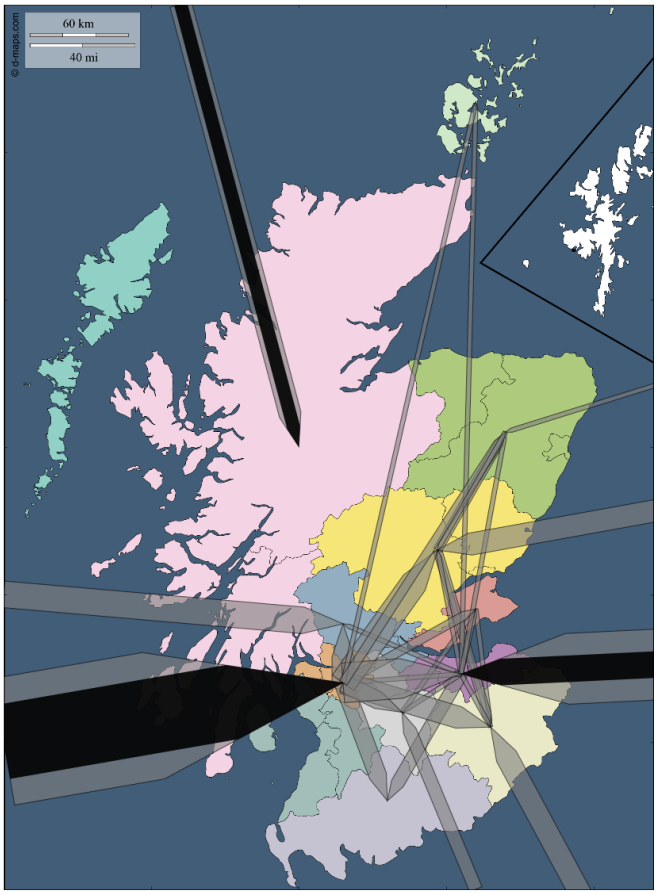


FIGURE 5.9: **Map showing inferred Victoria lineage movements based on HA.** Arrows indicate Yamagata lineage HA movements within Scotland. The grey transparent arrow indicates the upper 95% highest posterior density value for the number of virus movements along the path. The black arrow within the grey transparent arrow shows the lower 95% highest poterior density value, if it is absent it indicates that the lower 95% HPD interval includes zero. Arrows originating from outside of Scotland indicate inferred influenza B virus movements from the rest of the world into Scotland. There is strong evidence that Victoria lineage had been introduced independently into the Greater Glasgow and Clyde, Lothian and Highland NHS board areas with no strong support for Victoria lineage movement within Scotland. Map shown is from replicate 1.

latitudes have a tendency to be sinks for influenza viruses. Admittedly, the current data are severely lacking in sampling density and geographic extent, but raise some interesting questions. For example, even if sampling were improved across Scotland, how much could actually be learned about the movements of the virus? The estimated evolutionary rates for influenza B virus segments is between 1.2 and 2.2 $\times 10^{-3}$ substitutions per site per year. This naïvely corresponds to between 16.8 and 30.8 substitutions in a genome of nearly 14kb nucleotides each year or 1 substitution every 12 days to 3 weeks. Human travel usually takes place over much shorter timescales than this, so

Origin	Destination	Events/year	BF
Tayside	Grampian	1.15	4.23
Tayside	Borders	1.09	5.09
Lothian	Dumfries and Galloway	1.17	13.82
Lothian	Orkney	0.95	>30
Greater Glasgow and Clyde	Lanarkshire	1.66	>30
Greater Glasgow and Clyde	Fife	1.17	>30
Greater Glasgow and Clyde	Forth Valley	1.58	6.87
Greater Glasgow and Clyde	Dumfries and Galloway	1.43	>30
Greater Glasgow and Clyde	Tayside	1.99	>30
Greater Glasgow and Clyde	Lothian	2.40	>30
Greater Glasgow and Clyde	Grampian	1.17	>30
Greater Glasgow and Clyde	Borders	1.72	20.56
Greater Glasgow and Clyde	Orkney	0.74	3.20
Outside of Scotland	Forth Valley	0.60	>30
Outside of Scotland	Tayside	0.52	9.18
Outside of Scotland	Lothian	1.28	>30
Outside of Scotland	Greater Glasgow and Clyde	2.36	>30
Outside of Scotland	Highland	0.43	>30
Outside of Scotland	Borders	0.54	13.24

TABLE 5.3: **Summary of statistically significant (Bayes factor >3) NHS Board transitions for Victoria lineage.** Similar to Yamagata lineage, the inferred mean rate (events/year) is low for most transitions. In contrast to Yamagata lineage, however, more significant rates were found for Victoria lineage, including higher migration rates within Scotland. The cut off for statistical significance was considered to be Bayes factor (BF) above 3. Results are shown from replicate 1.

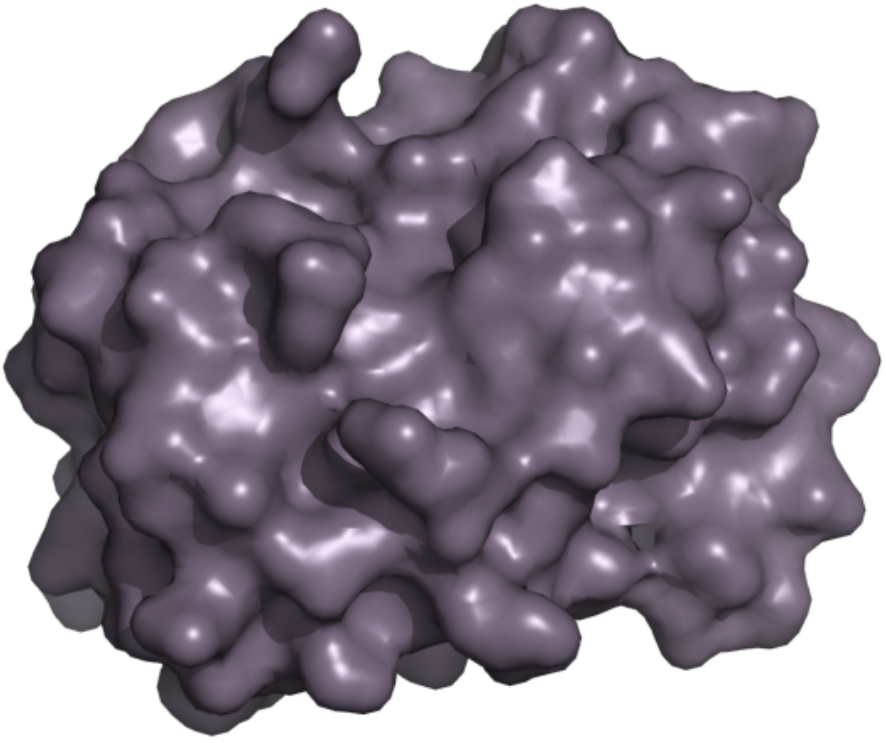
perhaps it is a bit unrealistic to expect to fully untangle the detailed migration history of pathogens from genetic data alone, especially given that full genome sequencing for pathogens like influenza viruses, although cheaper to do now using modern sequencing methods, is still not performed routinely.

Ethics statement

Handling and testing of specimens for the study was carried out in accordance with local ethical approval (South East Scotland SAHSC Human Annotated BioResource reference No.10/S1402/33).

CHAPTER 6

ORIGINS OF INFLUENZA D VIRUS



History was like some vast thing that was always over the tight horizon, invisible except in its effects. It was what happened when you weren't looking – an unknowable infinity of events, which although out of control, controlled everything.

– Kim Stanley Robinson, *Red Mars*

Introduction

Influenza D virus

Prior to 2011 the *Orthomyxoviridae* family of negative sense single stranded RNA viruses was known to contain 3 genera (Influenzaviruses A, B and C) that infect humans. The three genera have different host ranges: influenza A viruses (IAVs) infect a variety of vertebrate hosts: humans, birds, swine, seals, dogs, horses, ferrets and bats are currently known to be permissive to influenza A virus replication. Influenza B and C viruses (IBVs and ICVs, respectively), on the other hand, are thought to primarily infect humans with strong evidence that other species are permissive to their replication, but do not act as a reservoir (Guo et al., 1983; Osterhaus et al., 2000; Bodewes et al., 2013; Ran et al., 2015). All Orthomyxoviruses described since the discovery of influenza C virus in 1950 have belonged to clades distantly related to the clade containing influenza A, B and C viruses: ISA- (Krossøy et al., 1999), Quaranja- (Presti et al., 2009) and Thogotovirus (Leahy et al., 1997).

In 2013, however, a new member of *Orthomyxoviridae* belonging to the influenza virus clade was described after a nasal swab from a 15-week old pig with respiratory symptoms in Oklahoma was taken in April, 2011 (Hause et al., 2013). Enzymatic assays showed activity consistent with an O-acetylesterase, rather than a neuraminidase, suggesting a close affinity of this virus to influenza C viruses, although PCR and RT-PCR assays for influenza B and C viruses were negative. Following sequencing and *de novo* assembly it became clear that this virus is a distant relative of influenza C viruses and though initially suggested to be a divergent subtype of influenza C virus (Hause et al., 2013) it was later shown that this new virus exhibits a divergence from influenza C virus comparable to that between influenza A and B viruses and thus warrants the creation of a new genus within *Orthomyxoviridae*, influenzavirus D (Hause et al., 2014).

Despite being first isolated from a pig, influenza D virus (IDV) has been consistently isolated from cows ever since, first in USA (10 isolates from Kansas, Minnesota, Nebraska, Oklahoma and Texas) (Hause et al., 2013, 2014; Collin et al., 2015), then in China (3 isolates, all from Shandong) (Jiang et al., 2014) and most recently in France (1 isolate) (Ducatez et al., 2015). Phylogenetic incongruence among the 7 genomic

segments amongst USA isolates (Collin et al., 2015), *in vitro* experiments (Hause et al., 2014) and the data presented here (figure 6.5) argue for frequent reassortment in influenza D virus. In order for reassortment to be observed distinct lineages of the virus have to co-infect the same host and the same cell. Current estimates of influenza D virus prevalence vary wildly: when individual-level data are available, the prevalence in cattle ranges from 0.66% (China, Jiang et al. 2014) to 18% (USA, Hause et al. 2014) by rRT-PCR. Seroprevalence studies based on haemagglutination inhibition (HI) assays have suggested the presence of the virus in humans (1.3%, Hause et al. 2013) and relatively prevalent infection in swine (9.5%, Hause et al. 2013). HI assays performed on 141 serum samples from 8 cattle herds in the USA found that 7 herds had geometric mean titers to influenza D virus consistent with widespread infection (Hause et al., 2014), although information about individual titers is not available. However, seroprevalence studies can be unreliable, for example infections with highly pathogenic viruses, such as Ebola virus, have been suggested to be unrealistically prevalent in humans and chimpanzees (Leroy et al., 2004; Schoepp et al., 2014). Since rRT-PCR amplifies and quantifies the presence of RNA that is identifiably derived from influenza D virus, it is a more reliable method in detecting a specific virus, whereas antibody assays are at risk of detecting evidence of infection with related pathogens and/or can have a high false positive rate due to other factors, such as the age of the serum.

The recent discovery of an Orthomyxovirus in livestock, its high prevalence, wide geographic distribution and the distant possibility that the virus might be replication-competent in humans raises several questions. Some studies have suggested that cows might serve as a reservoir for influenza D virus, citing high prevalence of the virus and ease of virus isolation from cattle (Hause et al., 2014; Collin et al., 2015). However, it is difficult to establish whether cows are the "original" hosts of influenza D virus or a recent intermediate host that has played a role in the dissemination of the virus into swine due to the widespread raising of cattle in agriculture. To investigate whether cows are indeed the reservoir for influenza D virus or just the most recent successful host for the virus we use the available sequence data (14 genomes) to compare and contrast the population genetics of IDV with what is known about emerging infectious diseases and livestock pathogens.

Phylodynamics of RNA viruses

Adaptive immune responses of vertebrate hosts are capable of responding to viral challenge on the same timescales as virus evolution. These host defence mechanisms can shape viral diversity, for example influenza A viruses in humans have high rates of phenotypic change, called antigenic drift (Burnet, 1955), that allow the viruses to bypass existing population immunity. Usually the selective pressure on the virus is so high that a novel antigenic drift variant will frequently become the common ancestor of all future viruses circulating in humans, putting the entire virus genome through severe population bottlenecks (Buonagurio et al., 1986; Bedford et al., 2012). Other evolutionary forces, such as genetic drift can also push a virus population through a bottleneck, but in the presence of reassortment or recombination genetic drift must be very intense to result in a genome-wide bottleneck. In contrast, measles virus does not appear to change much antigenically over time (Rota et al., 1992) to the point that a vaccine developed from a strain isolated half a century ago is still effective today (Uzicanin and Zimmerman, 2011). Lack of antigenic change and lifelong immunity following infection limits measles virus infections to individuals that have never been exposed to the virus or the vaccine before, *i.e.* children. It has been shown that measles epidemics in unvaccinated children in the UK and France occur roughly every 2 years (Grenfell et al., 2001).

These contrasting degrees of access to susceptible hosts leave characteristic imprints on phylogenies of these viruses - rapid lineage turn over in influenza A virus results in spindly trees, where the most recent lineages are derived from a small number of dominant ancestral lineages with limited diversification at any given time point, usually sharing common ancestors within 5 or so years in the past. Because of lack of strong selection measles virus trees appear "bushier" (Grenfell et al., 2004), where recent lineages are derived from a diverse assemblage of ancestral lineages, which share common ancestors a number of years in the past. There are few viruses described with dynamics like those observed in human influenza A viruses and it has been suggested that most RNA virus lineages go extinct via genetic drift, *i.e.* turn over neutrally (Holmes, 2008). The degree to which viruses turn over neutrally should be further increased by hosts with short generation times, where reinfection with related viruses is expected to be rare. For example, influenza A viruses evolve at a similar rate in swine

and humans, but swine viruses undergo antigenic change at a considerably slower rate than human viruses (Jong et al., 2007). This is probably the result of swine being slaughtered around the age of 6 months. In addition, livestock get moved around via trade networks, putting viruses into contact with naïve host populations. As long as viruses are given access to a steady supply of naïve hosts, which would be expected from highly mobile pathogens, and the survival of viral lineages does not depend on host re-infection, the selective pressure on the virus would be expected to be minimal.

We seek to exploit this phenomenon in livestock viruses to address the origins of influenza D virus in the following way. Regardless of whether influenza D virus is a recent or old infection of cattle, there is little reason to suspect that this virus would be significantly different from other livestock viruses. The first assumption is that influenza D virus does not undergo severe population bottlenecks due to selection, but that stochastic extinction of lineages *i.e.* genetic drift can take place. Under a constant population size, coalescent theory states that the variance in expected time to common ancestor is high. Influenza viruses are segmented and in some cases, for example avian influenza A, human influenza B and influenza D virus, reassort often. Since reassortment unlinks each segment, each segment undergoes genetic drift largely independently. This leads to a simple prediction - long-term influenza virus-host associations will inevitably lead to variable times of most recent common ancestors (tMRCAs) in each segment, as diversity builds up over time, gets lost randomly from some segments through genetic drift or gets reassorted in from divergent lineages. The second assumption is that a zoonotic event will result in a perfect bottleneck, where effectively a single genome is introduced into a new host. If an influenza virus introduced into a new host species sufficiently recently, however, reassortment would only be able to reshuffle diversity accumulated since the population bottleneck at the time of zoonotic transmission and thus segment tMRCAs would be identical.

Methods

Sequence data

Genomic sequences of influenza D virus isolates from the USA (10 isolates) and China (3 isolates) were downloaded from GenBank. The genomic sequence of the French isolate was kindly provided by Dr. Mariette Ducatez (French National Institute for Agronomy Research). Only non-overlapping coding regions of each segment were kept following alignment with MUSCLE (Edgar, 2004). To find how clock-like the evolution of influenza D virus segments is maximum likelihood phylogenies were made in PhyML (Guindon and Gascuel, 2003) under a GTR+Γ_4 (Tavaré, 1986; Yang, 1994) nucleotide substitution model, mid-point rooted and sequence isolation dates regressed against tip divergence from the root.

Swine H3N2 and avian H7N7 data were downloaded from Influenza Research Database (Squires et al., 2012). Sequences were specifically chosen to be from viruses of the same subtype in order to avoid aligning the highly diverse HA and NA segments (Dugan et al., 2008). Exactly 14 genomes were sampled for both of these influenza A virus subtypes (one replicate each) and sequences were chosen roughly from a 3 year time window.

Simulated sequences were generated in fastsimcoal2 (Excoffier et al., 2013), under effective population sizes 5, 20 and 80. Simulations were set up with 3 demes. 14 genomes with 7 non-recombining "chromosomes" 2000 nucleotides long were simulated, with 10 genomes sampled from deme 1, 3 from deme 2 and 1 from deme 3. Mutation rate was arbitrarily set to 1×10^{-3} generation^{-1} site^{-1}, migration rates were uniform across demes and varied from 0.1 to 5.0 migrations per generation. Growth rates were varied from 0.0 (constant population size) to 5.0 per generation (according to the formula $N_t = N_0 e^{-rt}$, where t is generations).

Phylogenetic analyses

Phylogenetic analyses were carried out using BEAST (Drummond et al., 2012). Independent strict molecular clocks, as oppossed to relaxed molecular clocks (Drummond

et al., 2006), were specified for each segment in order to avoid over-parameterization and to accommodate apparent rate heterogeneity between segments. Marginal likelihood estimation-based model testing (Baele et al., 2012) was performed to identify the best combination of nucleotide substitution model partitioning and demographic model. Two independent HKY+Γ_4 (Hasegawa et al., 1985; Yang, 1994) models were specified for codon positions 1+2 and 3 and applied to either coding sequences in the entire genome jointly or to individual segments. Three parametric coalescent tree priors (constant population size, exponential growth and logistic growth) were tested. The demographic model tree prior was shared by all segments. All 8 models were run for 100 million states with the first 10 million discarded as burn-in. Marginal likelihood estimates based on both path-sampling and stepping stone sampling indicated the joint nucleotide substitution model, where parameters of the nucleotide substitution model are inferred for the entire genome, rather than individual segments, and logistic growth as the best combination of models for the data (table D.1). Maximum clade credibility trees were reconstructed in TreeAnnotator v1.8.2 using the common ancestor heights option (Heled and Bouckaert, 2013).

Results

Influenza D virus segments exhibit similar nucleotide site diversity

We observe differences in nucleotide site diversity (calculated as mean pairwise differences between sequences) across the single sets of 14 genomes of swine H3N2 and avian H7N7 influenza A viruses sampled over a similar time period (≈3 years) to influenza D virus (figure 6.1). This conforms to the prediction from coalescent theory that under a constant population size and in the presence of reassortment or recombination the variance in time to most recent common ancestor across the genome exhibits high variance, since genetic drift happens at sub-genomic level. Influenza D virus, on the other hand, displays little variance in nucleotide site diversity across all codon positions (figure 6.2), as well as low overall genetic diversity (figure 6.1).

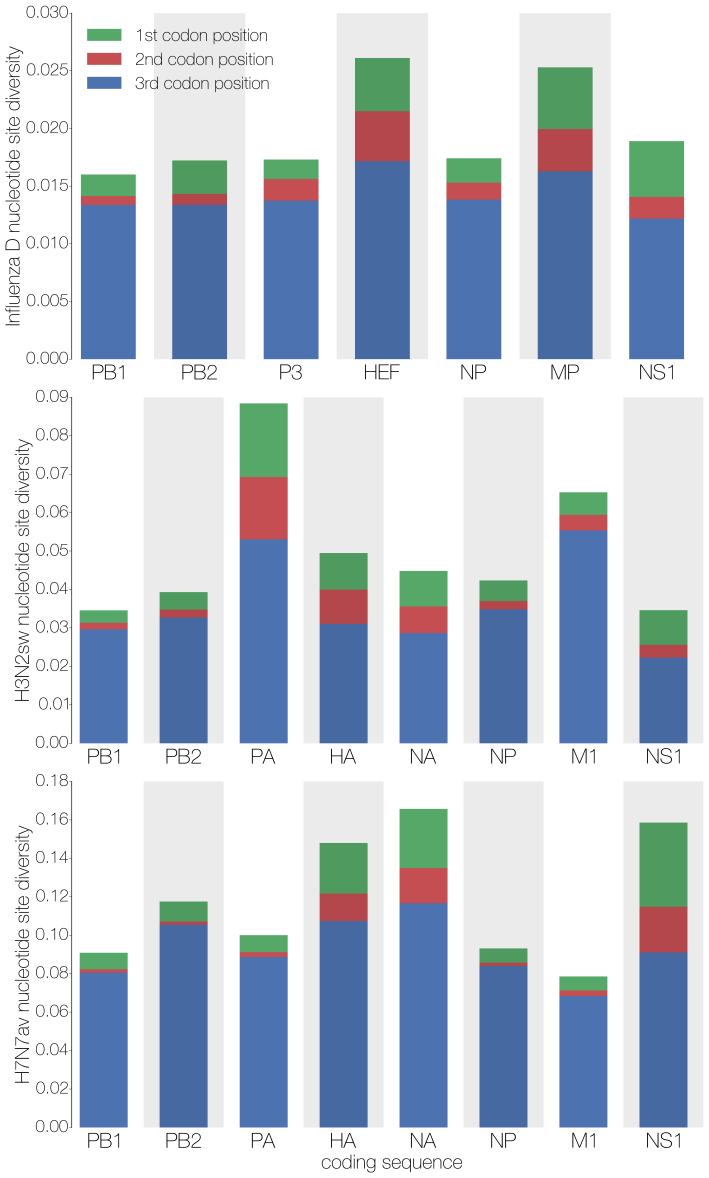


FIGURE 6.1: **Nucleotide site diversity in influenza D, A/H3N2sw and A/H7N7av segments.** Influenza D virus exhibits low and consistent nucleotide site diversity across all genomic segments. Genomes of influenza A viruses (swine H3N2 and avian H7N7) on the other hand are heterogenous with respect to nucleotide site diversity, suggestive of gene flow and/or genetic drift in some but not all segments. Note that y axes are different for all three plots and that influenza D virus has lower overall nucleotide site diversity than A/H3N2sw and both have lower nucleotide site diversity than A/H7N7av.

Clock-like evolution of most influenza D virus segments

Despite small numbers of influenza D virus sequences and their limited temporal sampling range (the oldest and the most recent strains were isolated about 3 years apart), some coding sequences display a reasonable correlation between virus isolation dates

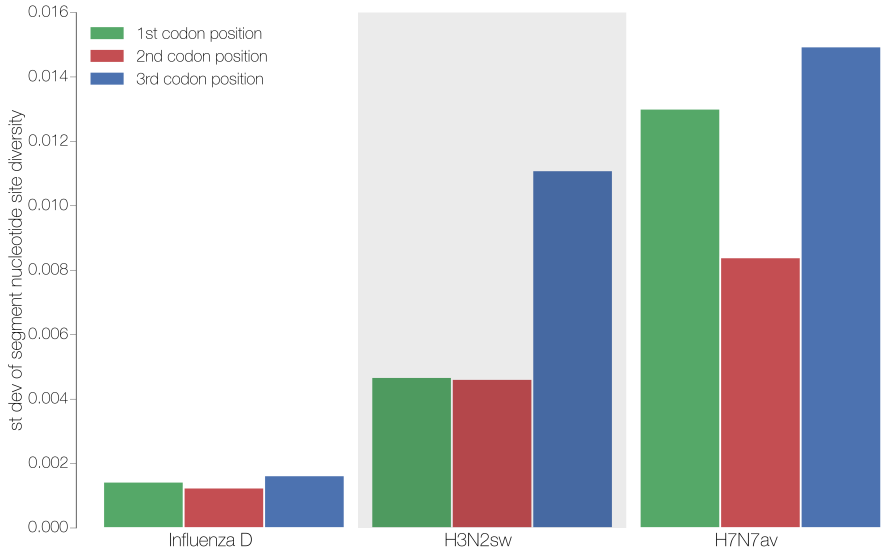


FIGURE 6.2: **Standard deviation of segment nucleotide site diversities of influenza D, A/H3N2sw and A/H7N7av viruses.** Influenza D virus segments have low deviation in nucleotide site diversity, consistent across all codon positions. This is not the case for influenza A viruses (swine H3N2 and avian H7N7), where a lot of variation exists in nucleotide site diversity between segments and codon positions.

and sequence divergence from the root of the tree (figure 6.3). The slopes of these regressions are entirely consistent with previously published estimates of RNA virus evolutionary rates (Jenkins et al., 2002), on the order of 1×10^{-3} substitutions per site per year. Bayesian estimates of the rate are highly consistent as well, with fairly narrow confidence intervals (figure 6.4). HEF, the surface glycoprotein of influenza C-like Orthomyxoviruses, exhibits a higher substitution rate compared to other segments. This has been reported previously for many other viruses, including influenza A viruses (Hayashida et al., 1985).

We also confirm the presence of reassortment reported previously (Hause et al., 2014; Collin et al., 2015). Time calibrated phylogenies of influenza D virus segments are significantly incongruent, for example US isolates form a well-supported monophyletic clade in PB1, but Chinese isolate diversity in P3 falls within the diversity of US isolates, with good support (figure 6.5). For most segments it is difficult to recover a well supported branching order near the root.

Logistic growth was the preferred choice of demographic model (Baele et al., 2012), although different demographic tree priors yielded very similar results. Most of the

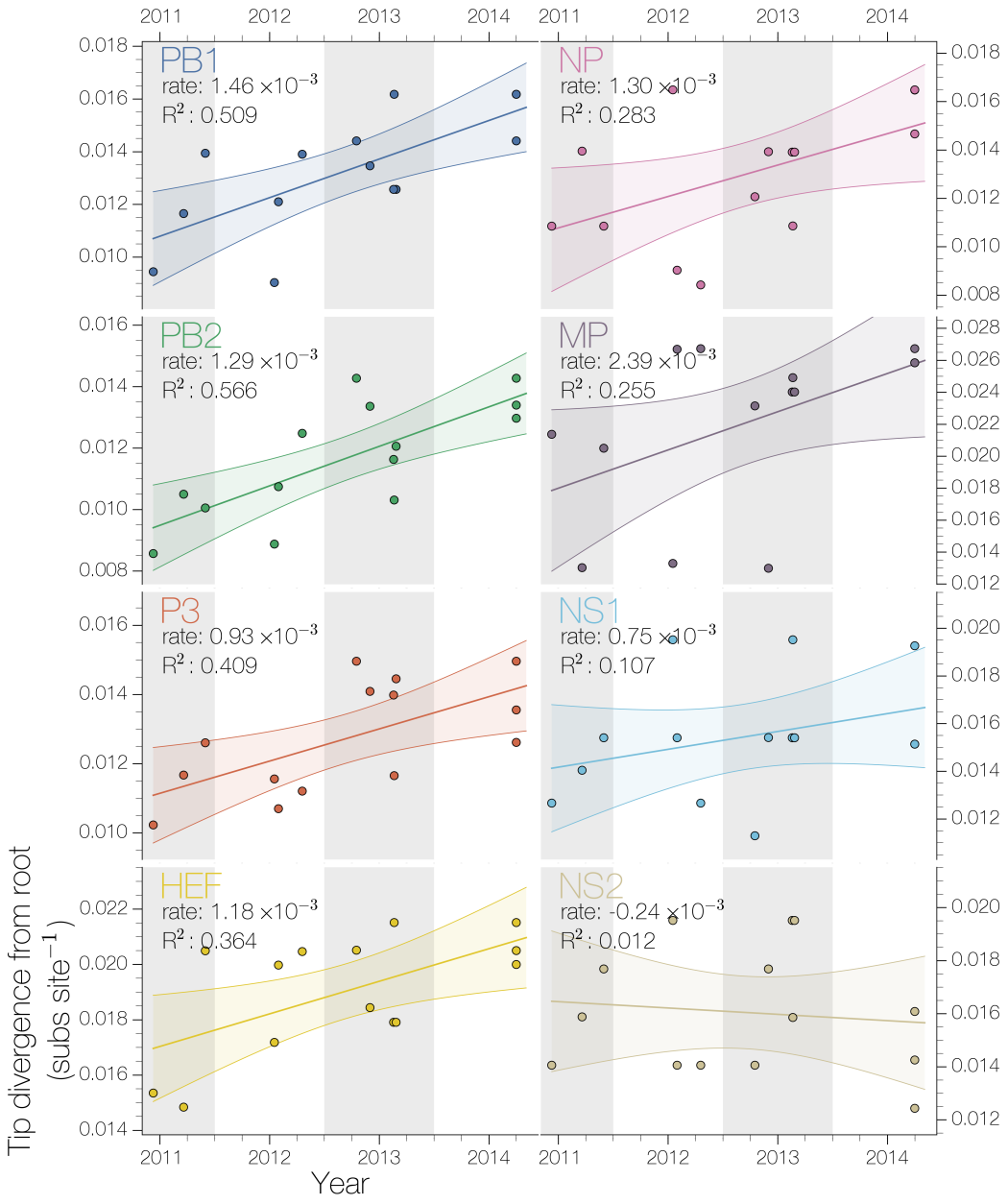


FIGURE 6.3: **Root to tip regressions for influenza D virus coding sequences.** All coding sequences except for NS2 exhibit a positive correlation between isolation date and divergence from the root, consistent with a molecular clock. Inferred rates of evolution (slope) are consistent with rates of evolution described for many other RNA viruses.

probability density of the combined posterior distributions of segment tMRCAs is concentrated around 2001.16, with 95% highest posterior density intervals 1994.92–2006.65, although variation exists in both mean and HPDs of tMRCA distributions within any given segment (figure 6.6). In addition, most segment phylogenies exhibit a large degree of population structure. For example isolates from the USA and China appear to be

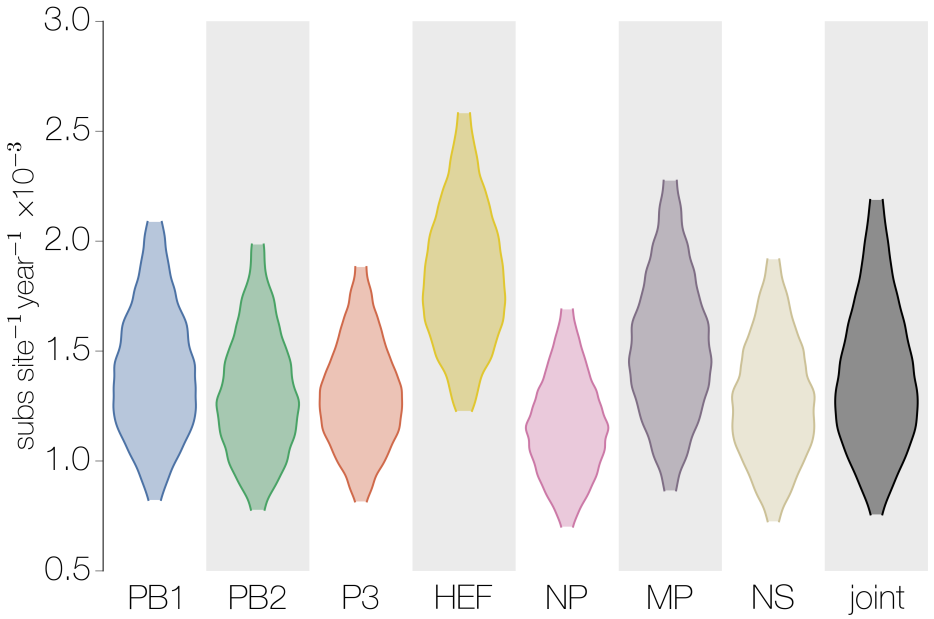


FIGURE 6.4: **Segment evolutionary rate estimates.** Violin plots of evolutionary rate estimates for each segment and a genome wide estimate from combined MCMC samples. The width of each violin indicates the normalized kernel density estimate of samples from the posterior distribution. The HEF segment has an evolutionary rate higher than other segments, suggesting some degree of selection for presumably immune escape.

more closely related to isolates from the same country in all segments, although Chinese isolates are probably epidemiologically linked. Similarly, the French lineage does not appear to have reassorted in its immediate past, suggesting lack of contact with other lineages.

Global spread of influenza D virus

Influenza D virus, as a member of the influenza virus branch of the *Orthomyxoviridae*, is unlikely to be spread via an arthropod vector like its distant relatives – Thogoto- (Haig et al., 1965; Anderson and Casals, 1973) and Quaranjaviruses (Kemp et al., 1975; Clifford et al., 1968). The isolation of the virus on three distant continents would suggest that influenza D virus has likely spread via cattle trade networks. However, alternative vehicles for viral movement, such as vertebrates that migrate (birds, bats, humans) or are traded (swine, sheep) cannot be discounted. Analysis of cumulative cattle trade between the years 2000 and 2011 (data from the Food and Agriculture Organization Corporate Statistical Database, FAOSTAT, available at `faostat.fao.org/` and downloaded on March 26, 2015) suggests that cattle trade is mostly localized

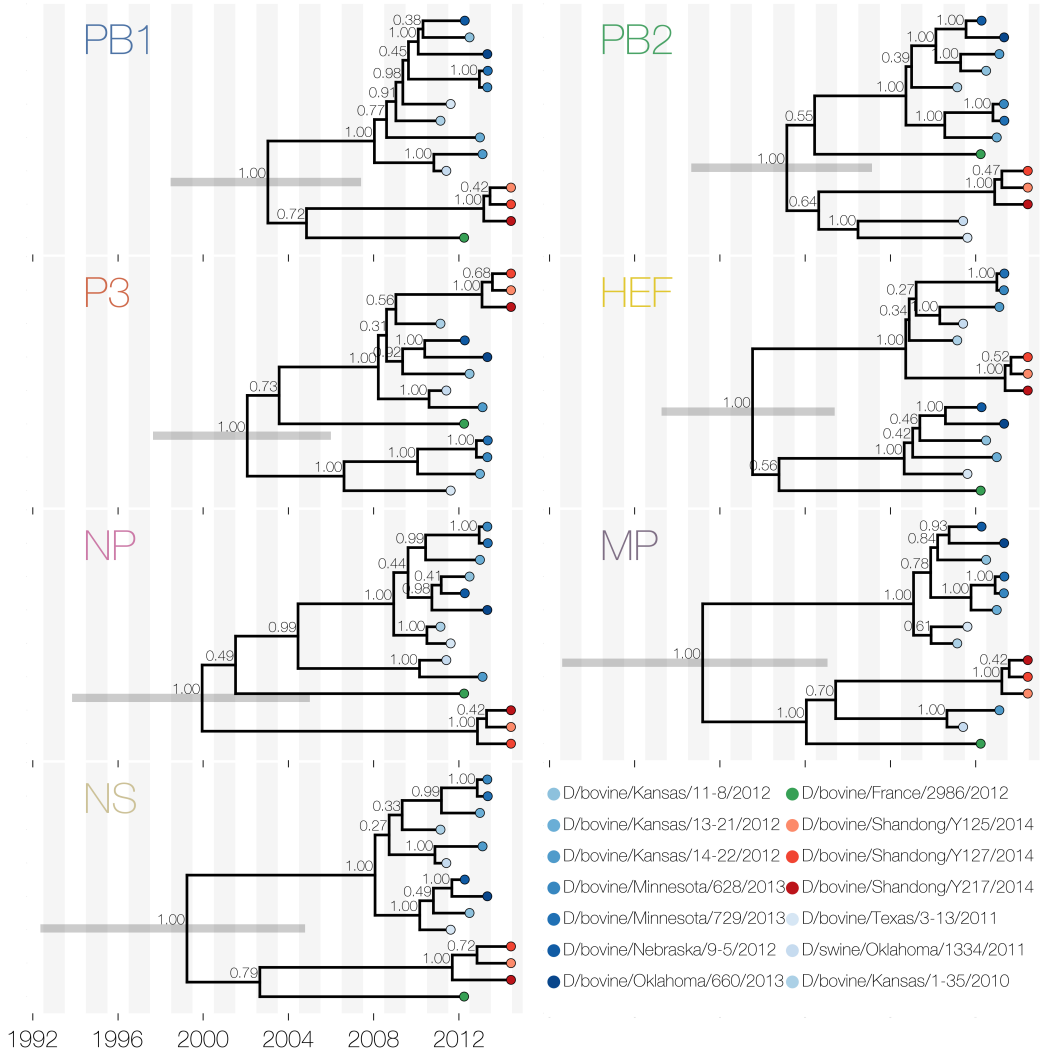


FIGURE 6.5: **Phylogenetic trees of influenza D virus genomic segments.** Maximum clade credibility trees for each influenza D virus segment recovered under the logistic growth demographic prior. Branches are coloured by clade posterior probability and tips are coloured by country of isolation and isolate. The grey horizontal block at the root of each tree is the 95% HPD interval for the time of most recent common ancestor in that segment.

- highest numbers of cattle traded are between countries that are on the same continent (figure 6.7). In fact, none of the countries where influenza D virus has been isolated (France, China and USA) have traded live cattle to a significant degree with each other: the highest number of head exported between them is from the US to China, around 2700 head over 11 years. France, China and USA do not share significant cattle trade partners either. Canada, Indonesia and Thailand have exported cattle to all countries reporting the presence of influenza D virus, albeit at limited numbers. Canada, which is a major trade partner of the USA with nearly 12 million head cattle exported to the

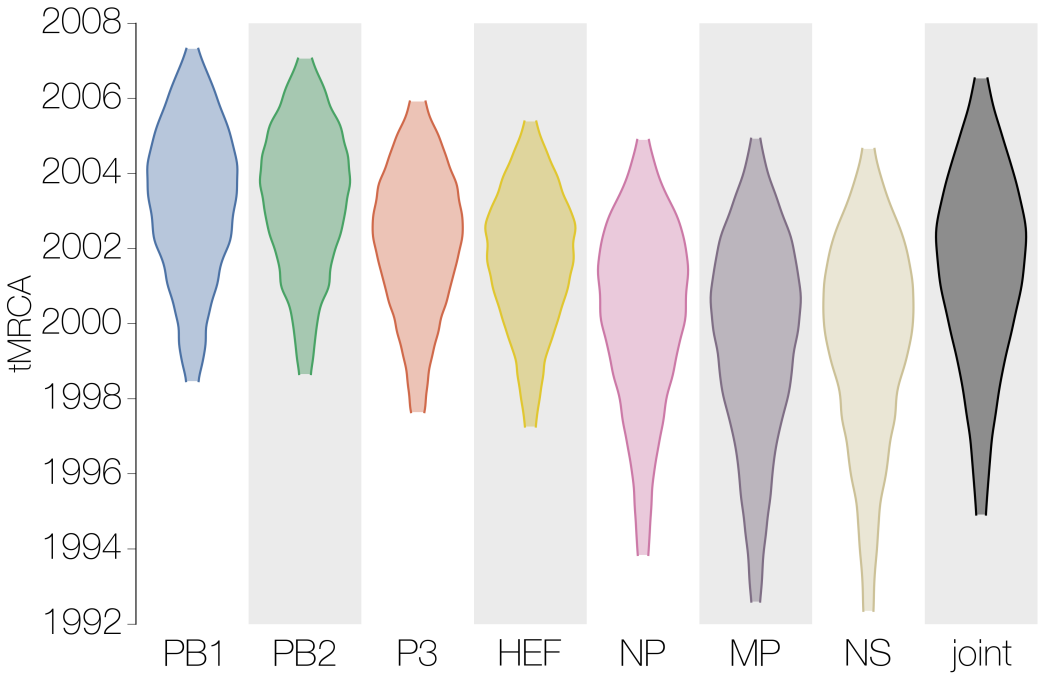


FIGURE 6.6: **Segment tMRCA estimates.** Violin plots of the estimated date of the root for trees shown in figure 6.5.

US between 2000 and 2011, has only exported around 1500 head to China over the same time period and only 40 head to France. Indonesia has exported 357 head to France and negligible numbers to USA and China, whereas Thailand has exported 281 head to the US between 2000 and 2011, but much smaller numbers to China and France. Overall, these results argue that cattle trade is unlikely to be responsible for the global spread of the virus, but could be an important contributor to regional dispersal.

High migration rates and low population sizes result in lower diversity

Coalescent simulations indicate that mean nucleotide site diversity across 7 non-recombining segments is highest when the overall effective population size is large, migration rates are low and population size is not experiencing rapid growth (figure 6.8). This is an intuitive result from coalescent theory – genetic drift is slower in large populations, especially so if they are constant over time. The coalescence rate in growing populations is rapid initially and slows down as the population size increases. Structured populations, on the other hand, by their nature do not allow lineages to coalesce until they are put into contact with each other via migration. In the absence of any migration lineages will never coalesce, but with increasing migration rate structured populations

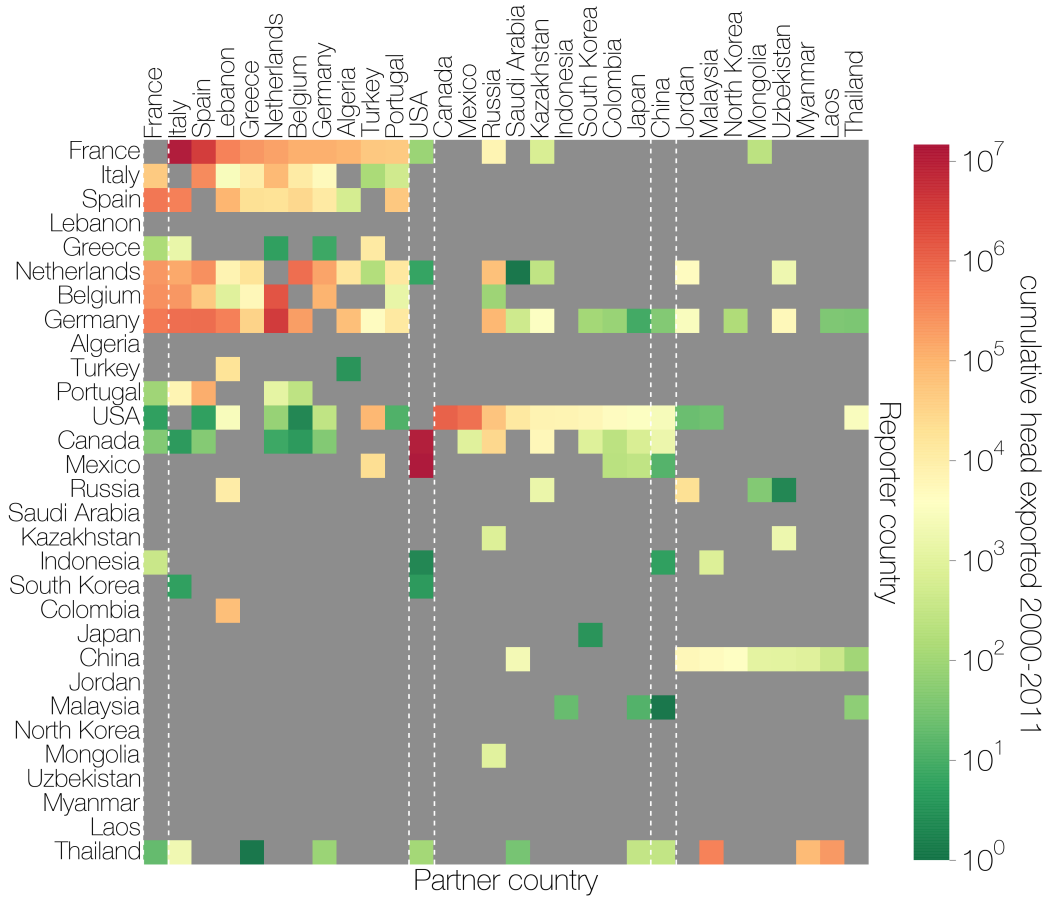


FIGURE 6.7: **Cattle trade network of top 10 partners of countries reporting influenza D virus.** Matrix shows the reported cumulative live cattle exports of USA, France and China and their top 10 cattle trade partners by number of head exported. Most cattle are moved to neighbouring countries. Only Canada, Thailand and Indonesia have reported exporting cattle to all three affected countries.

become more panmictic. Under small effective population sizes, migration rate is the dominant force affecting genetic diversity, rather than population growth rate. This is because contact between demes is necessary for coalescence to occur, so even though lineages coalesce rapidly within demes due to their small effective population sizes, they require migration to allow coalescence to occur between demes.

Broadly similar patterns can be described for the variance in nucleotide site diversity across segments (figure 6.8). Like nucleotide site diversity, high variance in nucleotide site diversity across segments is observed when migration rate is low, effective population size is high and the population size is constant. This is because the rate of coalescence in individual segments is affected by the effective population size – a population bottleneck will increase the rate of genetic drift in all segments, making it very

likely that segments will coalesce at similar times. It is because of this that lower variance in nucleotide site diversity exists under high population growth rates, when the population size is small in the past it forces all segments to coalesce rapidly. The effect of migration rate on variance in nucleotide site diversity across segments arises from the fact that migrants may carry different lineages of each segment and under low migration rates it takes numerous migrations to bring all of the segment lineages into the same deme in order for them to coalesce.

Discussion

Origin of influenza D virus

Our results suggest that influenza D virus is likely to be a new and recent introduction of an Orthomyxovirus into cattle. Swine, as a potential reservoir for influenza viruses or segment lineages with zoonotic potential, have been under surveillance for a long time, yet there is no evidence of influenza D virus infection in swine prior to 2011. We believe that cryptic circulation of this virus in cattle, which are rarely tested for Orthomyxoviruses, is consistent with its recent isolation and description. The sequence data available now for influenza D virus, though limited, can be used to address the initial questions about the origins of the virus.

We base our conclusion on the consistency of nucleotide site diversities and by extension estimated dates of tMRCAs across influenza D virus segments (figures 6.1 and 6.5). Coalescent theory states that under a constant population size the variance in tMRCA is high and proportional to the effective population size. In the presence of strong population structure (low migration between demes) this variance is even greater, as lineages have to wait for migration before they can coalesce. We show this prediction, which is consistent with coalescent theory, using simulated data (figure 6.8). Given the phylogenies (figure 6.5) we recovered, influenza D virus appears to have very low migration rates. If the virus were seeded to the 3 affected countries from a virus population that has been stable over time we would expect to see considerable variation in segment tMRCAs, especially in light of its apparently low rate of migration. Under expontential population growth the variance in nucleotide site diversity is minimal

(figure 6.8) and this is what we observe in influenza D virus, leading us to conclude that influenza D virus effective population size was much lower in the past, most consistent with an introduction of a single genomic lineage of the virus. Exponential population growth would be expected to take place following an introduction of an antigenically novel pathogen into a naïve host population, although we have no evidence that this has been the case for influenza D virus due to limited data.

Much less can be said about the geographic origin of influenza D virus. Influenza D virus had sufficient time to accumulate genetic diversity since its proposed introduction, without any of its diversity being sampled prior to 2011 (figure 6.5). The long period of time of unsampled diversity greatly reduces the statistical power of phylogeographic models to infer the location of the root, as lineages could have switched location several times during their unsampled history. Even if the virus had been detected soon after its introduction the limited geographic sampling would preclude a robust phylogeographic analysis. Cattle trade network data do not appear to be helpful either – no significant direct cattle trade links between France, China and USA exist and there are even fewer cattle being moved via shared trade partners (figure 6.7). This suggests that other vectors for the virus might be responsible for spreading the virus globally, and could include other livestock, wild migratory vertebrates or even humans.

Predictions and future directions

In the absence of sequence data some predictions can be made based entirely on the phylogenetic position of influenza D virus in the *Orthomyxoviridae* family. If the original host of influenza D virus is ever identified it will most likely be a mammal. All influenza viruses (A, B, C and D) isolated to date infect mammals, with the exception of influenza A viruses, which is the only genus in the family known to have achieved sustained transmission in birds. Having said that, identifying original hosts of viruses causing acute disease is extremely difficult. This is because at any given time only a small portion of the host population has actively replicating virus. Previously it took extremely wide sampling to tentatively identify fruit bats as a potential reservoir for Ebola virus (Leroy et al., 2005).

Although the available data for influenza D virus is limited to 14 sequenced isolates and limited serological studies, the analyses presented here lead to two testable predictions. Assuming that sequencing efforts so far have captured all of the existing influenza D virus diversity since its introduction, and if it were possible to reliably test cattle sera collected prior to 1995 there should be no evidence of influenza D virus infection. Likewise, we predict the presence of antibodies against the virus in sera collected past 2007. These predictions are based on the combined estimates of segment tMRCAs (mean 2001.16, 95% HPDs: 1994.92–2006.65, figure 6.4). Finally, given the sheer volume of cattle moved between the affected countries and their neighbours, especially in the case of USA and France, we expect the virus to be widespread in Europe and North America.

Note - there have been two studies published since the analyses presented in this chapter. One of the studies has shown the presence of influenza D viruses in Mississippi cattle closely related to other US influenza D virus isolates (Ferguson et al., 2015). The other study has shown that influenza D virus is present in swine and cattle in Italy (Chiapponi et al., 2016). The latter result is of particular interest, since Italian isolates are closely related to US isolates, rather than the French isolate as was expected from the analyses presented here. Although this highlights how little is known about this novel pathogen at present, every additional genome of influenza D virus from a new location has the potential to dramatically advance our understanding of the origins and spread of this enigmatic virus.

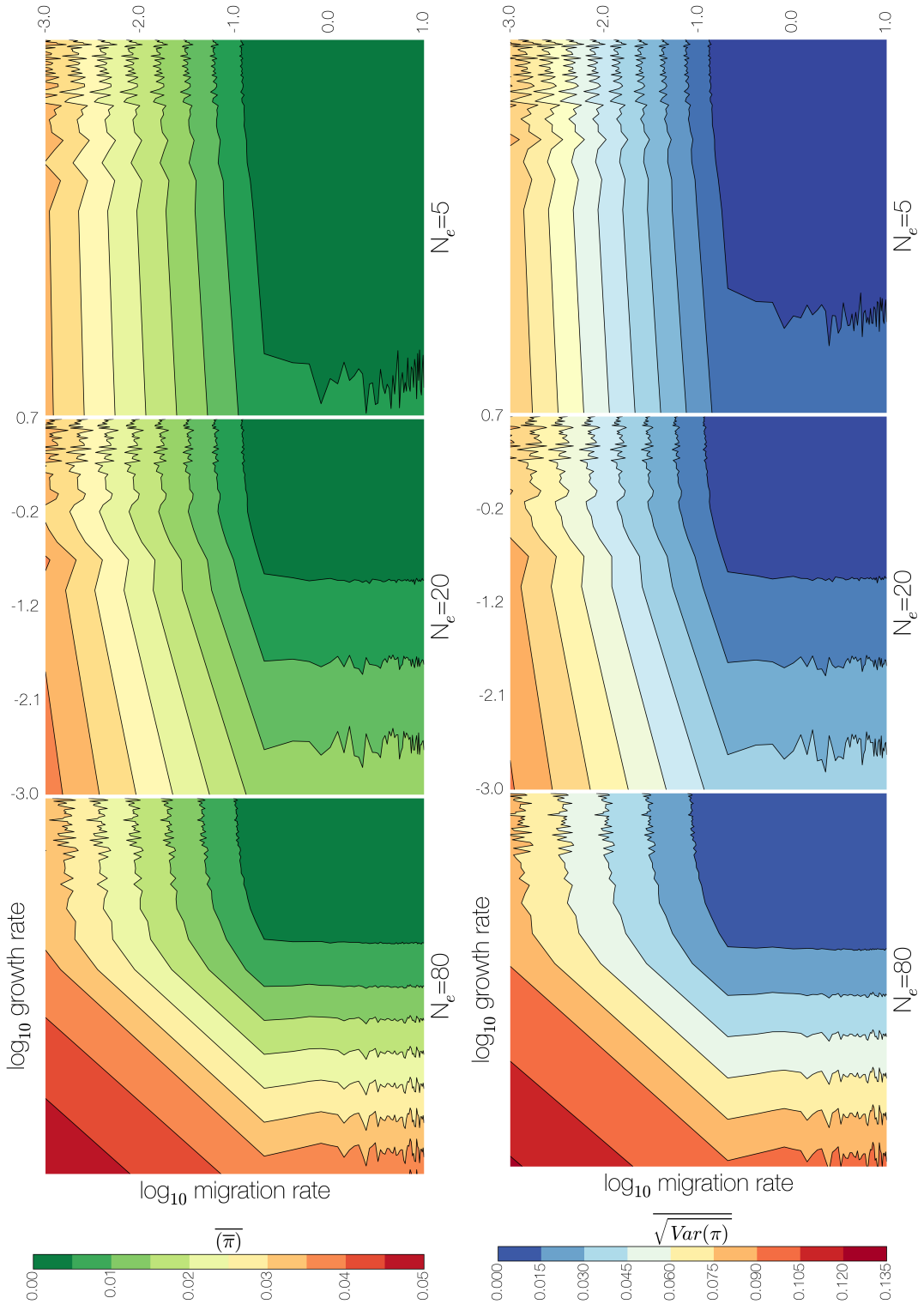
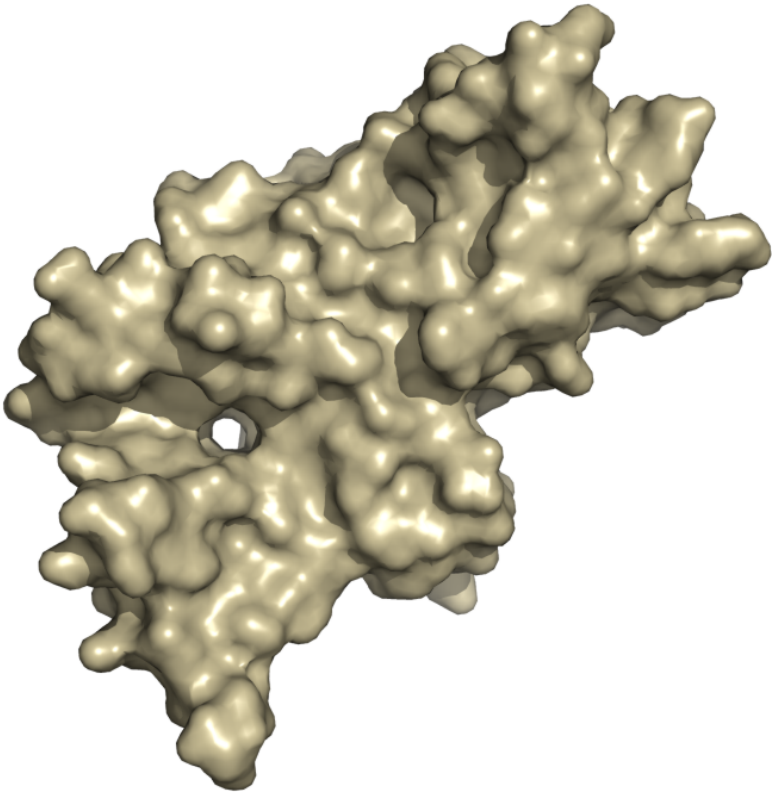


FIGURE 6.8: **Mean nucleotide site diversity (left) and mean standard deviation of nucleotide site diversity (right) under varied effective population sizes, migration rates and growth rates.** Contour plots indicate mean nucleotide site diversity (left) and mean standard deviation in nucleotide site diversity (right) between 7 "segments" under 3 different effective population sizes (5, 20 and 80, top to bottom), varying migration rates (x axes) and varying growth rates (y axes) (2500 parameter combinations) across 100 replicate simulations. Genetic diversity accumulates and the highest variance in nucleotide site diversity across segments is observed under large effective population sizes, reduced migration and constant population size.

CHAPTER 7

CONCLUDING REMARKS



But just because something does not have an ending doesn't mean it doesn't have a conclusion.

– Iain M. Banks, *Use of Weapons*

A unifying thread that runs through most of this thesis is reticulate evolution of RNA viruses. It is easy to account for, and indeed to comprehensively investigate, in segmented negative sense RNA viruses, since they do not recombine within segments often. When it comes to other groups of viruses which are known to recombine, many studies choose to overlook recombination altogether or account for it in rudimentary ways. Nevertheless, reticulate evolution is not always a nuisance parameter to be integrated out during analysis, as has been shown in some of the work presented here. RNA viruses do not recombine obligately in the same way sexual organisms do, so the presence of any form of reticulate evolution is usually indicative of an appreciable degree of co-infection. The mere presence of reticulate evolution in RNA viruses thus provides information which is otherwise inaccesible to regular phylogenetic analyses save for catching distinct lineages in the act of co-infection.

Chapter 2 explores the selection regimes experienced by Ebola virus strain Makona following nearly 2 years of circulation in humans in West Africa. It shows that despite the magnitude and geographic extent of the outbreak, the virus behind it is entirely conforming to expectations from population genetics theory. I show that the most successful lineages of EBOV in West Africa have fewer non-synonymous changes than lineages observed only once in the data, consistent with transient circulation of mildly deleterious genotypes. Purifying selection across the entire genome of the Makona strain is relaxed compared to long-term selective pressures, as one would predict from the epidemiology of the virus in humans and from the way lineages from the reservoir get sampled.

In chapter 3 I have shown that MERS coronavirus genomes exhibit signals consistent with recombination. The finding in itself is not particularly surprising, since it is a virus from a taxonomic group that is known to be recombinogenic. Placed within an epidemiological context, however, recombination tells us that there is considerable genetic diversity of the virus to detect recombination in the first place and more importantly, that diverse lineages encounter each other inside hosts at sufficient frequency to yield recombinant lineages. Furthermore, virtually nothing is known of the evolutionary dynamics of the MERS coronavirus in dromedary camels, the presumed reservoir of the virus. This complements serological studies, which have indicated widespread infection with MERS-CoV among camels, using virus sequences mostly isolated from humans.

Chapter 4 deals with reticulate evolution in human influenza B viruses since their split into two genomic lineages in humans. The two lineages have reassorted infrequently over their 30 years of co-circulation, but sufficiently often to homogenize the diversity in some segments. In fact of all 16 lineages that have existed initially (two lineages across eight segments), five went extinct following reassortment. Only three segments of currently circulating influenza B viruses still display the 30 year-old split into two lineages: PB1, PB2 and HA. More interestingly, these three segments also display a large degree of co-divergence and all successful influenza B viruses have maintained PB1, PB2 and HA segment complexes composed entirely of Victoria or Yamagata lineage and though the association between PB1 and PB2 segments could intuitively be explained due to the intimate proximity with which the proteins operate in the polymerase heterotrimer, the association between polymerase segments and HA remains entirely mysterious. Chapter 4 thus describes evidence of molecular-level constraint within influenza B virus genomes inferred entirely from patterns of reticulate evolution in genetic data.

Slightly more ambitious than others, chapter 5 is an attempt at elucidating both within lineage reassortment dynamics of influenza B viruses and their geographic diffusion within a confined geographic region. Although sequence sampling is far from ideal, especially to address the latter objective, the chapter highlights a problem that will only get worse as sequencing capacity improves. Inference of both within-lineage reassortment and migration depend on evolutionary rates being high enough to "capture" the processes of interest. Within-lineage reassortment necessarily involves lineages that are closely related and thus more difficult to infer reassortment from, in contrast to highly divergent lineages described in chapter 4. Likewise, human travel duration is much shorter than the time it takes for an influenza virus genome to acquire a substitution that will be sampled by surveillance.

Chapter 6 attempts to make meaningful inferences about the origins of influenza D virus from very limited data. By heavily relying on coalescent theory to fill in the gaps left by late and incomplete sampling I propose that influenza D virus is an infection that has been introduced into cattle populations relatively recently. This is based on the inherent variance of the coalescent process. This variance is proportional to the effective population size and most importantly the independence of segment genealogies. I show that despite short sampling periods there is sufficient molecular clock signal,

which is further amplified by having segments share a demographic history. The results suggest that all currently sampled influenza D virus segment lineages share a common ancestor within the last 20 years, which is consistent with an introduction of a single virus genome, followed by diversification. Although not formally tested, I note that the influenza D virus population appears to be highly structured geographically. All isolates are more closely related to isolates from the same country, even in quite diverse clades. How the virus ended up on three separate continents within 20 years, however, is an entirely different question. More genome sequences of this pathogen and from many locations around the world will become available in the future, which will shed more light on the origins and global spread of influenza D virus.

Chapters 3, 4, 5, 6 in this thesis thus cover several important aspects of RNA virus evolution that can be investigated by exploiting the lack of non-clonal descent, which are: cryptic circulation (chapter 3), speciation (chapter 4), short-term genome dynamics (chapter 5) and emergence in new hosts (chapter 6). In particular, a relatively high number of non-recombining segments in influenza viruses and their fast evolutionary rate collectively make these members of *Orthomyxoviridae* one of the most tractable systems available for investigating the effects of reticulate evolution on genomic organization, interactions between segments and population dynamics.

Although not the major focus of the thesis, the inference of RNA virus dynamics in reservoir hosts from sequences sampled in novel hosts is also touched upon in chapters 2 and 3. Both Ebola virus and MERS coronavirus have emerged in humans several times and even though sequences of both viruses are available from their presumed natural hosts, most of the sequence data come from affected humans. In the case of Ebola virus in chapter 2, the inference of population dynamics in the natural reservoir from viruses sequenced in humans is limited. It is clear that viruses in the natural reservoir undergo typical population dynamics with genetic drift determining lineage turnover, purifying selection maintaining protein function and some evidence of positive selection in regions of the genome expected to be targeted by the adaptive immune system. Much more can be said about MERS coronavirus dynamics in the reservoir from the analyses of sequences mostly sampled from humans which were presented in chapter 3. This is partly because both the reservoir and humans have been sampled well, and partly because MERS-CoV emerges in humans relatively often. Due to the presence

of recombination, which can only occur during co-infection, and the relative rarity of human MERS infection the analyses presented in chapter 3 indicated that the natural hosts of MERS-CoV must be co-infected at a high frequency.

One of the more compelling ideas that has evolved as a result of the work presented in this thesis is the use of software in innovative ways. This could not have been possible if the software employed throughout the thesis were black boxes – impossible to investigate when something goes wrong and providing answers without showing the maths. BEAST gets a lot of praise from me for that. Very few results presented in this thesis can be arrived at directly from basic analyses in BEAST, instead relying on teasing apart information from posterior samples after BEAST has done the heavy lifting. This carries a two-fold benefit - not only does it allow a newcomer some transparency, it also provides much more flexibility to the types of analyses that a person could do. Although the development of highly generalised and transparent software is not necessarily a given, I would personally like for that to be the case in the future.

Other advances in the field are easier to foresee. Quantifying within-patient diversity of viruses and integrating such data with modern phylogenetic methods is inevitable. This will allow for vast improvements in statistical power to infer the direction of transmission in epidemiological scenarios, but also to better understand the evolutionary dynamics of viruses at a much finer scale. Most intriguing of all is the prospect of being able to quantify Darwinian selection the way it has been defined for decades, that is changes in allele frequencies, rather than looking at the kinds of mutations that have accumulated in different parts of the virus genome. Methods to detect and reconstruct non-clonal evolution are also likely to evolve. In the special case of RNA viruses, where time can be an essential source of information, these methods are likely to advance the fastest, since only lineages that co-exist can recombine or reassort. Finally, continuing innovations in sequencing technologies will expand our understanding of virus evolution at all scales, from the aforementioned quantification of within-patient diversity to the billions of years of evolution that have shaped virus genomes and their taxonomic groups we see today. By now I actually suspect that within a few years when sequencing technologies are sufficiently advanced we will find that RNA viruses, the fastest evolving protein coding organisms on the planet, are evolving too slow to keep pace with our curiosity and patience.

BIBLIOGRAPHY

Aburizaiza AS, Mattes FM, Azhar EI, Hassan AM, Memish ZA, Muth D, Meyer B, Lattwein E, Müller M, Drosten C. 2013. Investigation of anti-MERS-coronavirus antibodies in blood donors and abbatoir workers in Jeddah and Makkah, Kingdom of Saudi Arabia, fall 2012. *Journal of Infectious Diseases*. p. jit589.

Adney DR, van Doremalen N, Brown VR, Bushmaker T, Scott D, de Wit E, Bowen RA, Munster VJ. 2014. Replication and shedding of MERS-CoV in upper respiratory tract of inoculated dromedary camels. *Emerging Infectious Diseases*. 20:1999–2005.

Aiewsakun P, Katzourakis A. 2015. Endogenous viruses: Connecting recent and ancient viral evolution. *Virology*. 479–480:26–37.

Anderson CR, Casals J. 1973. Dhori virus, a new agent isolated from Hyalomma dromedarii in India. *The Indian Journal of Medical Research*. 61:1416–1420.

Anisimova M, Nielsen R, Yang Z. 2003. Effect of Recombination on the Accuracy of the Likelihood Method for Detecting Positive Selection at Amino Acid Sites. *Genetics*. 164:1229–1236.

Ansaldi F, D'Agaro P, de Florentiis D, et al. (12 co-authors). 2003. Molecular characterization of influenza B viruses circulating in northern italy during the 2001–2002 epidemic season. *Journal of Medical Virology*. 70:463–469.

Ascenzi P, Bocedi A, Heptonstall J, Capobianchi MR, Di Caro A, Mastrangelo E, Bolognesi M, Ippolito G. 2008. Ebolavirus and Marburgvirus: Insight the Filoviridae family. *Molecular Aspects of Medicine*. 29:151–185.

Assiri A, Al-Tawfiq JA, Al-Rabeeah AA, et al. (13 co-authors). 2013. Epidemiological, demographic, and clinical characteristics of 47 cases of middle east respiratory syndrome coronavirus disease from saudi arabia: a descriptive study. *The Lancet Infectious Diseases*. 13:752–761.

Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution*. 29:2157–2167.

Baize S, Pannetier D, Oestereich L, et al. (31 co-authors). 2014. Emergence of Zaire Ebola Virus Disease in Guinea. *New England Journal of Medicine*. 371:1418–1425.

Baltimore D. 1971. Expression of animal virus genomes. *Bacteriological Reviews*. 35:235–241.

Baric RS, Shieh CK, Stohlman SA, Lai MMC. 1987. Analysis of intracellular small RNAs of mouse hepatitis virus: evidence for discontinuous transcription. *Virology*. 156:342–354.

Basler CF, Wang X, Mühlberger E, Volchkov V, Paragas J, Klenk HD, García-Sastre A, Palese P. 2000. The Ebola virus VP35 protein functions as a type I IFN antagonist. *Proceedings of the National Academy of Sciences*. 97:12289–12294.

Bedford T, Cobey S, Beerli P, Pascual M. 2010. Global migration dynamics underlie evolution and persistence of human influenza A (H3N2). *PLoS Pathog*. 6:e1000918.

Bedford T, Cobey S, Pascual M. 2011. Strength and tempo of selection revealed in viral gene genealogies. *BMC Evolutionary Biology*. 11:220. PMID: 21787390.

Bedford T, Rambaut A, Pascual M. 2012. Canalization of the evolutionary trajectory of the human influenza virus. *BMC Biology*. 10:38.

Bedford T, Riley S, Barr IG, et al. (26 co-authors). 2015. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*. 523:217–220.

Bedford T, Suchard MA, Lemey P, Dudas G, Gregory V, Hay AJ, McCauley JW, Russell CA, Smith DJ, Rambaut A. 2014. Integrating influenza antigenic dynamics with molecular evolution. *eLife*. 3.

Beemon K, Duesberg P, Vogt P. 1974. Evidence for Crossing-Over Between Avian Tumor Viruses Based on Analysis of Viral RNAs. *Proceedings of the National Academy of Sciences of the United States of America*. 71:4254–4258.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. *Nucleic Acids Research*. 41:D36–D42.

Bergeron C, Valette M, Lina B, Ottmann M. 2010. Genetic content of influenza H3N2 vaccine seeds. *PLoS Currents*. 2:RRN1165.

Bergh O, Børsheim KY, Bratbak G, Heldal M. 1989. High abundance of viruses found in aquatic environments. *Nature*. 340:467–468.

Biek R, Pybus OG, Lloyd-Smith JO, Didelot X. 2015. Measurably evolving pathogens in the genomic era. *Trends in Ecology & Evolution*. 30:306–313.

Bielejec F, Lemey P, Carvalho LM, Baele G, Rambaut A, Suchard MA. 2014. πBUSS: a parallel BEAST/BEAGLE utility for sequence simulation under complex evolutionary scenarios. *BMC Bioinformatics*. 15:133.

Bodewes R, Morick D, de Mutsert G, et al. (11 co-authors). 2013. Recurring influenza B virus infections in seals. *Emerging Infectious Diseases*. 19:511–512.

Bogner P, Capua I, Lipman DJ, Cox NJ, et al. (5 co-authors). 2006. A global initiative on sharing avian flu data. *Nature*. 442:981–981.

Boni MF, de Jong MD, van Doorn HR, Holmes EC. 2010. Guidelines for Identifying Homologous Recombination Events in Influenza A Virus. *PLoS ONE*. 5:e10434.

Boni MF, Posada D, Feldman MW. 2007. An Exact Nonparametric Method for Inferring Mosaic Structure in Sequence Triplets. *Genetics*. 176:1035–1047.

Boni MF, Zhou Y, Taubenberger JK, Holmes EC. 2008. Homologous recombination is very rare or absent in human influenza A virus. *Journal of Virology*. 82:4807–4811. PMID: 18353939.

Boucher CA, Cammack N, Schipper P, Schuurman R, Rouse P, Wainberg MA, Cameron JM. 1993. High-level resistance to (-) enantiomeric 2'-deoxy-3'-thiacytidine in vitro is due to one amino acid substitution in the catalytic site of human immunodeficiency virus type 1 reverse transcriptase. *Antimicrobial Agents and Chemotherapy*. 37:2231–2234.

Broberg E, Beauté J, Snacken R. 2013. Fortnightly influenza surveillance review, 9th May. Technical report, European Centre for Disease Prevention and Control, Stockholm.

Bruen TC, Philippe H, Bryant D. 2006. A Simple and Robust Statistical Test for Detecting the Presence of Recombination. *Genetics*. 172:2665–2681.

Buonagurio DA, Nakada S, Parvin JD, Krystal M, Palese P, Fitch WM. 1986. Evolution of human influenza A viruses over 50 years: rapid, uniform rate of change in NS gene. *Science*. 232:980–982.

Burnet SFM. 1955. Principles of animal virology. Academic Press.

Carpenter JA, Keegan LP, Wilfert L, O'Connell MA, Jiggins FM. 2009. Evidence for ADAR-induced hypermutation of the Drosophila sigma virus (Rhabdoviridae). *BMC Genetics*. 10:75.

Carroll SA, Towner JS, Sealy TK, McMullan LK, Khristova ML, Burt FJ, Swanepoel R, Rollin PE, Nichol ST. 2013. Molecular Evolution of Viruses of the Family Filoviridae Based on 97 Whole-Genome Sequences. *Journal of Virology*. 87:2608–2616.

Cauchemez S, Fraser C, Van Kerkhove MD, Donnelly CA, Riley S, Rambaut A, Enouf V, van der Werf S, Ferguson NM. 2014. Middle east respiratory syndrome coronavirus: quantification of the extent of the epidemic, surveillance biases, and transmissibility. *The Lancet Infectious Diseases*. 14:50–56.

Chang SF, Sgro JY, Parrish CR. 1992. Multiple amino acids in the capsid structure of canine parvovirus coordinately determine the canine host range and specific antigenic and hemagglutination properties. *Journal of Virology*. 66:6858–6867.

Chare ER, Gould EA, Holmes EC. 2003. Phylogenetic analysis reveals a low rate of homologous recombination in negative-sense RNA viruses. *Journal of General Virology*. 84:2691–2703.

Chen R, Holmes EC. 2008. The evolutionary dynamics of human influenza B virus. *Journal of Molecular Evolution*. 66:655–663.

Chiapponi C, Faccini S, De Mattia A, Baioni L, Barbieri I, Rosignoli C, Nigrelli A, Foni E. 2016. Detection of Influenza D Virus among Swine and Cattle, Italy. *Emerging Infectious Disease journal*. 22.

Chu DK, Poon LL, Gomaa MM, et al. (13 co-authors). 2014. MERS coronaviruses in dromedary camels, egypt. *Emerging Infectious Diseases*. 20:1049–1053.

Clifford CM, Thomas LA, Hughes LE, Kohls GM, Philip CB. 1968. Identification and Comparison of two Viruses Isolated from Ticks of the Genus Ornithodoros. *The American Journal of Tropical Medicine and Hygiene*. 17:881–885.

Cobbin JCA, Verity EE, Gilbertson BP, Rockman SP, Brown LE. 2013. The source of the PB1 gene in influenza vaccine reassortants selectively alters the hemagglutinin content of the resulting seed virus. *Journal of Virology*. 87:5577–5585. PMID: 23468502.

Collin EA, Sheng Z, Lang Y, Ma W, Hause BM, Li F. 2015. Cocirculation of two distinct genetic and antigenic lineages of proposed influenza D virus in cattle. *Journal of Virology*. 89:1036–1042.

Corman VM, Ithete NL, Richards LR, Schoeman MC, Preiser W, Drosten C, Drexler JF. 2014a. Rooting the phylogenetic tree of Middle East Respiratory Syndrome Coronavirus by characterization of a conspecific virus from an African bat. *Journal of Virology*. 88:11297–11303.

Corman VM, Jores J, Meyer B, et al. (13 co-authors). 2014b. Antibodies against MERS coronavirus in dromedary camels, Kenya, 1992–2013. *Emerging Infectious Diseases*. 20.

Cotten M, Watson SJ, Kellam P, et al. (22 co-authors). 2013. Transmission and evolution of the Middle East Respiratory Syndrome Coronavirus in Saudi Arabia: a descriptive genomic study. *The Lancet*. 382:1993–2002.

Cotten M, Watson SJ, Zumla AI, et al. (20 co-authors). 2014. Spread, circulation, and evolution of the Middle East Respiratory Syndrome Coronavirus. *mBio*. 5:e01062–13.

de Oliveira T, Pybus OG, Rambaut A, et al. (14 co-authors). 2006. Molecular Epidemiology: HIV-1 and HCV sequences from Libyan outbreak. *Nature*. 444:836–837.

Didelot X, Falush D. 2007. Inference of bacterial microevolution using multilocus sequence data. *Genetics*. 175:1251–1266.

Dobzhansky T. 1937. Genetics and the Origin of Species. Columbia University Press.

Dong G, Luo J, Zhang H, Wang C, Duan M, Deliberto TJ, Nolte DL, Ji G, He H. 2011. Phylogenetic diversity and genotypical complexity of H9N2 influenza a viruses revealed by genomic sequence analysis. *PLoS ONE*. 6:e17212.

Drosten C, Muth D, Corman V, et al. (15 co-authors). 2014. An observational, laboratory-based study of outbreaks of MERS-coronavirus in Jeddah and Riyadh, Kingdom of Saudi Arabia, 2014. *Clinical Infectious Diseases*. p. ciu812.

Drosten C, Seilmaier M, Corman VM, et al. (22 co-authors). 2013. Clinical features and virological analysis of a case of Middle East Respiratory Syndrome Coronavirus infection. *The Lancet Infectious Diseases*. 13:745–751.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol*. 4:e88.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Molecular Biology and Evolution*. 22:1185–1192.

Drummond AJ, Suchard MA. 2010. Bayesian random local clocks, or one rate to rule them all. *BMC Biology*. 8:114.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*. 29:1969–1973.

Ducatez MF, Pelletier C, Meyer G. 2015. Influenza D Virus in Cattle, France, 2011–2014. *Emerging Infectious Diseases*. 21.

Dudas G, Bedford T, Lycett S, Rambaut A. 2015. Reassortment between Influenza B Lineages and the Emergence of a Coadapted PB1–PB2–HA Gene Complex. *Molecular biology and evolution*. 32:162–172.

Dudas G, Rambaut A. 2014. Phylogenetic Analysis of Guinea 2014 EBOV Ebolavirus Outbreak. *PLoS Currents*. 6.

Duffy S, Shackelton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics*. 9:267–276.

Dugan VG, Chen R, Spiro DJ, et al. (15 co-authors). 2008. The Evolutionary Genetics and Emergence of Avian Influenza Viruses in Wild Birds. *PLoS Pathogens*. 4.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*. 32:1792–1797.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust demographic inference from genomic and SNP data. *PLoS Genet*. 9:e1003905.

Felsenstein J. 1973. Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. *Systematic Biology*. 22:240–249.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*. 17:368–376.

Felsenstein J. 1985a. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*. 39:783–791.

Felsenstein J. 1985b. Phylogenies and the Comparative Method. *The American Naturalist*. 125:1–15.

Ferguson L, Eckard L, Epperson WB, Long LP, Smith D, Huston C, Genova S, Webby R, Wan XF. 2015. Influenza D virus infection in Mississippi beef cattle. *Virology*. 486:28–34.

Ferreira MAR, Suchard MA. 2008. Bayesian analysis of elapsed times in continuous-time markov chains. *Canadian Journal of Statistics*. 36:355–368.

Fitch WM, Leiter JM, Li XQ, Palese P. 1991. Positive Darwinian evolution in human influenza A viruses. *Proceedings of the National Academy of Sciences*. 88:4270–4274.

Fitch WM, Margoliash E. 1967. Construction of Phylogenetic Trees. *Science*. 155:279–284.

Fodor E, Smith M. 2004. The PA subunit is required for efficient nuclear accumulation of the PB1 subunit of the influenza a virus RNA polymerase complex. *Journal of Virology*. 78:9144–9153. PMID: 15308710.

Francis T. 1940. A New Type of Virus from Epidemic Influenza. *Science*. 92:405–408.

Fulvini AA, Ramanunninair M, Le J, Pokorny BA, Arroyo JM, Silverman J, Devis R, Bucher D. 2011. Gene constellation of influenza A virus reassortants with high growth phenotype prepared as seed candidates for vaccine production. *PLoS ONE*. 6:e20823.

Gierer S, Hofmann-Winkler H, Albuali WH, et al. (11 co-authors). 2013. Lack of MERS Coronavirus neutralizing antibodies in humans, Eastern Province, Saudi Arabia. *Emerging Infectious Diseases*. 19:2034–2036.

Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. 2013. Improving bayesian population dynamics inference: A coalescent-based model for multiple loci. *Molecular Biology and Evolution*. 30:713–724.

Gire SK, Goba A, Andersen KG, et al. (58 co-authors). 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*. 345:1369–1372.

Gog JR, Afonso EDS, Dalton RM, Leclercq I, Tiley L, Elton D, Kirchbach JCv, Naffakh N, Escriou N, Digard P. 2007. Codon conservation in the influenza A virus genome defines RNA packaging signals. *Nucleic Acids Research*. 35:1897–1907.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*. 11:725–736.

Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ. 2006. Nidovirales: Evolving the largest RNA virus genome. *Virus Research*. 117:17–37.

Gregory V, Bennett M, Orkhan M, Hajjar SA, Varsano N, Mendelson E, Zambon M, Ellis J, Hay A, Lin Y. 2002. Emergence of influenza A H1N2 reassortant viruses in the human population during 2001. *Virology*. 300:1 – 7.

Grenfell BT, Bjørnstad ON, Kappey J. 2001. Travelling waves and spatial hierarchies in measles epidemics. *Nature*. 414:716–723.

Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, Holmes EC. 2004. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science*. 303:327–332.

Guan Y, Zheng BJ, He YQ, et al. (18 co-authors). 2003. Isolation and Characterization of Viruses Related to the SARS Coronavirus from Animals in Southern China. *Science*. 302:276–278.

Gubareva LV, Kaiser L, Matrosovich MN, Soo-Hoo Y, Hayden FG. 2001. Selection of influenza virus mutants in experimentally infected volunteers treated with oseltamivir. *Journal of Infectious Diseases*. 183:523–531.

Guilligay D, Tarendeau F, Resa-Infante P, et al. (11 co-authors). 2008. The structural basis for cap binding by influenza virus polymerase subunit PB2. *Nature Structural & Molecular Biology*. 15:500–506.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*. 52:696–704.

Guo YJ, Jin FG, Wang P, Wang M, Zhu JM. 1983. Isolation of influenza C virus from pigs and experimental infection of pigs with influenza C virus. *The Journal of General Virology*. 64 (Pt 1):177–182.

Haig DA, Woodall JP, Danskin D. 1965. Thogoto Virus: a Hitherto Undescribed Agent Isolated from Ticks in Kenya. *Journal of General Microbiology*. 38:389–394.

Hall MD, Knowles NJ, Wadsworth J, Rambaut A, Woolhouse MEJ. 2013. Reconstructing Geographical Movements and Host Species Transitions of Foot-and-Mouth Disease Virus Serotype SAT 2. *mBio*. 4:e00591–13.

Han GZ, Boni MF, Li SS. 2010. No observed effect of homologous recombination on influenza C virus evolution. *Virology Journal*. 7:227. PMID: 20840780.

Hasegawa M, Kishino H, Yano Ta. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*. 22:160–174.

Hause BM, Collin EA, Liu R, Huang B, Sheng Z, Lu W, Wang D, Nelson EA, Li F. 2014. Characterization of a Novel Influenza Virus in Cattle and Swine: Proposal for a New Genus in the Orthomyxoviridae Family. *mBio*. 5:e00031–14.

Hause BM, Ducatez M, Collin EA, et al. (13 co-authors). 2013. Isolation of a novel swine influenza virus from Oklahoma in 2011 which is distantly related to human influenza C viruses. *PLoS pathogens*. 9:e1003176.

Hay AJ, Gregory V, Douglas AR, Lin YP. 2001. The evolution of human influenza viruses. *Philosophical Transactions of the Royal Society of London. Series B*. 356:1861–1870. PMID: 11779385 PMCID: PMC1088562.

Hayashida H, Toh H, Kikuno R, Miyata T. 1985. Evolution of influenza virus genes. *Molecular Biology and Evolution*. 2:289–303.

Hedrick PW, Thomson G. 1986. A two-locus neutrality test: Applications to humans, E. coli and lodgepole pine. *Genetics*. 112:135–156. PMID: 3510942.

Heled J, Bouckaert RR. 2013. Looking for trees in the forest: summary tree from posterior samples. *BMC Evolutionary Biology*. 13:221.

Heled J, Drummond AJ. 2008. Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology*. 8:289.

Herrewegh AAPM, Smeenk I, Horzinek MC, Rottier PJM, Groot RJd. 1998. Feline coronavirus type II strains 79-1683 and 79-1146 originate from a double recombination between feline coronavirus type i and canine coronavirus. *Journal of Virology*. 72:4508–4514.

Hill WG, Robertson A. 1968. Linkage disequilibrium in finite populations. *Theoretical and applied genetics*. 38:226–231.

Hiromoto Y, Saito T, Lindstrom SE, Li Y, Nerome R, Sugita S, Shinjoh M, Nerome K. 2000. Phylogenetic analysis of the three polymerase genes (PB1, PB2 and PA) of influenza B virus. *Journal of General Virology*. 81:929–937. PMID: 10725418.

Ho SYW. 2009. An examination of phylogenetic models of substitution rate variation among lineages. *Biology Letters*. 5:421–424.

Ho SYW, Lanfear R, Bromham L, Phillips MJ, Soubrier J, Rodrigo AG, Cooper A. 2011. Time-dependent rates of molecular evolution. *Molecular Ecology*. 20:3087–3101.

Ho SYW, Phillips MJ, Cooper A, Drummond AJ. 2005. Time Dependency of Molecular Rate Estimates and Systematic Overestimation of Recent Divergence Times. *Molecular Biology and Evolution*. 22:1561–1568.

Ho SYW, Shapiro B. 2011. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Molecular Ecology Resources*. 11:423–434.

Hoenen T, Safronetz D, Groseth A, et al. (24 co-authors). 2015a. Mutation rate and genotype variation of Ebola virus from Mali case sequences. *Science*. 348:117–119.

Hoenen T, Safronetz D, Groseth A, et al. (24 co-authors). 2015b. Erratum for the Report "Mutation rate and genotype variation of Ebola virus from Mali case sequences". *Science*. 348:aac5674.

Hoffmann E, Mahmood K, Yang CF, Webster RG, Greenberg HB, Kemble G. 2002. Rescue of influenza B virus from eight plasmids. *Proceedings of the National Academy of Sciences*. 99:11411–11416. PMID: 12172012.

Holmes EC. 2008. Evolutionary History and Phylogeography of Human Viruses. *Annual Review of Microbiology*. 62:307–328.

Holmes EC, Ghedin E, Miller N, et al. (11 co-authors). 2005. Whole-Genome Analysis of Human Influenza A Virus Reveals Multiple Persistent Lineages and Reassortment among Recent H3N2 Viruses. *PLoS Biol*. 3:e300.

Holmes EC, Urwin R, Maiden MC. 1999. The influence of recombination on the population structure and evolution of the human pathogen neisseria meningitidis. *Molecular Biology and Evolution*. 16:741–749.

Hon CC, Lam TY, Shi ZL, Drummond AJ, Yip CW, Zeng F, Lam PY, Leung FCC. 2008. Evidence of the Recombinant Origin of a Bat Severe Acute Respiratory Syndrome (SARS)-Like Coronavirus and Its Implications on the Direct Ancestor of SARS Coronavirus. *Journal of Virology*. 82:1819–1826.

Hoover K, Grove M, Gardner M, Hughes DP, McNeil J, Slavicek J. 2011. A Gene for an Extended Phenotype. *Science*. 333:1401–1401.

Ito H, Watanabe S, Takada A, Kawaoka Y. 2001. Ebola Virus Glycoprotein: Proteolytic Processing, Acylation, Cell Tropism, and Detection of Neutralizing Antibodies. *Journal of Virology*. 75:1576–1580.

Jenkins GM, Rambaut A, Pybus OG, Holmes EC. 2002. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *Journal of Molecular Evolution*. 54:156–165.

Jetzt AE, Yu H, Klarmann GJ, Ron Y, Preston BD, Dougherty JP. 2000. High Rate of Recombination throughout the Human Immunodeficiency Virus Type 1 Genome. *Journal of Virology*. 74:1234–1240.

Jiang WM, Wang SC, Peng C, Yu JM, Zhuang QY, Hou GY, Liu S, Li JP, Chen JM. 2014. Identification of a potential novel type of influenza virus in Bovine in China. *Virus Genes*. 49:493–496.

Johnson KM, Lange JV, Webb PA, Murphy FA. 1977. Isolaton and partial characterisation of a new virus causing acute haemorrhagic fever in Zaire. *The Lancet*. 309:569–571.

Jong JCd, Smith DJ, Lapedes AS, et al. (11 co-authors). 2007. Antigenic and Genetic Evolution of Swine Influenza A (H3N2) Viruses in Europe. *Journal of Virology*. 81:4315–4322.

Jukes T, Cantor C. 1969. Evolution of protein molecules. In: Munro M, editor, Mammalian protein metabolism, Academic Press, volume III, pp. 21–132.

Kanegae Y, Sugita S, Endo A, Ishida M, Senya S, Osako K, Nerome K, Oya A. 1990. Evolutionary pattern of the hemagglutinin gene of influenza B viruses isolated in Japan: cocirculating lineages in the same epidemic season. *Journal of Virology*. 64:2860–2865.

Kawaoka Y, Krauss S, Webster RG. 1989. Avian-to-human transmission of the PB1 gene of influenza A viruses in the 1957 and 1968 pandemics. *Journal of Virology*. 63:4603–4608. PMID: 2795713 PMCID: PMC251093.

Keck JG, Matsushima GK, Makino S, Fleming JO, Vannier DM, Stohlman SA, Lai MM. 1988. In vivo RNA-RNA recombination of coronavirus in mouse brain. *Journal of Virology*. 62:1810–1813.

Kemp GE, Lee VH, Moore DL. 1975. Isolation of Nyamanini and Quaranfil Viruses from Argas (Persicargas) Arboreus Ticks in Nigeria. *Journal of Medical Entomology.* 12:535–537.

Kendal AP, Noble GR, Skehel JJ, Dowdle WR. 1978. Antigenic similarity of influenza A(H1N1) viruses from epidemics in 1977–1978 to "Scandinavian" strains isolated in epidemics of 1950–1951. *Virology.* 89:632–636.

King AMQ, Adams MJ, Lefkowitz EJ. 2011. Virus Taxonomy: Classification and Nomenclature of Viruses : Ninth Report of the International Committee on Taxonomy of Viruses. Elsevier.

Kingman JFC. 1982. On the Genealogy of Large Populations. *Journal of Applied Probability.* 19:27–43.

Kirkegaard K, Baltimore D. 1986. The mechanism of RNA recombination in poliovirus. *Cell.* 47:433–443.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. 2006. Automated phylogenetic detection of recombination using a genetic algorithm. *Molecular Biology and Evolution.* 23:1891–1901.

Kottier SA, Cavanagh D, Britton P. 1995. Experimental evidence of recombination in coronavirus infectious bronchitis virus. *Virology.* 213:569–580.

Krossøy B, Hordvik I, Nilsen F, Nylund A, Endresen C. 1999. The Putative Polymerase Sequence of Infectious Salmon Anemia Virus Suggests a New Genus within the Orthomyxoviridae. *Journal of Virology.* 73:2136–2142.

Ksiazek TG, Erdman D, Goldsmith CS, et al. (26 co-authors). 2003. A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. *New England Journal of Medicine.* 348:1953–1966.

Lahti DC, Johnson NA, Ajie BC, Otto SP, Hendry AP, Blumstein DT, Coss RG, Donohue K, Foster SA. 2009. Relaxed selection in the wild. *Trends in Ecology & Evolution.* 24:487–496.

Lai MM, Baric RS, Makino S, Keck JG, Egbert J, Leibowitz JL, Stohlman SA. 1985. Recombination between nonsegmented RNA genomes of murine coronaviruses. *Journal of Virology*. 56:449–456.

Lam TTY, Zhu H, Chong YL, Holmes EC, Guan Y. 2015. Puzzling origins of the Ebola outbreak in the Democratic Republic of the Congo, 2014. *Journal of Virology*. pp. JVI.01226–15.

Lau SKP, Li KSM, Huang Y, et al. (14 co-authors). 2010. Ecoepidemiology and Complete Genome Comparison of Different Strains of Severe Acute Respiratory Syndrome-Related Rhinolophus Bat Coronavirus in China Reveal Bats as a Reservoir for Acute, Self-Limiting Infection That Allows Recombination Events. *Journal of Virology*. 84:2808–2819.

Le Guenno B, Formenty P, Formentry P, Wyers M, Gounon P, Walker F, Boesch C. 1995. Isolation and partial characterisation of a new strain of Ebola virus. *Lancet*. 345:1271–1274.

Leahy MB, Dessens JT, Weber F, Kochs G, Nuttall PA. 1997. The fourth genus in the Orthomyxoviridae: sequence analyses of two Thogoto virus polymerase proteins and comparison with influenza viruses. *Virus Research*. 50:215–224.

Lee JE, Fusco ML, Hessell AJ, Oswald WB, Burton DR, Saphire EO. 2008. Structure of the Ebola virus glycoprotein bound to an antibody from a human survivor. *Nature*. 454:177–182.

Lee JE, Saphire EO. 2009. Neutralizing ebolavirus: structural insights into the envelope glycoprotein and antibodies targeted against it. *Current Opinion in Structural Biology*. 19:408–417.

Lemey P, Minin VN, Bielejec F, Pond SLK, Suchard MA. 2012. A counting renaissance: combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection. *Bioinformatics*. 28:3248–3256.

Lemey P, Rambaut A, Drummond AJ, Suchard MA. 2009a. Bayesian Phylogeography Finds Its Roots. *PLoS Comput Biol*. 5:e1000520.

Lemey P, Suchard M, Rambaut A. 2009b. Reconstructing the initial global spread of a human influenza pandemic. *PLoS Currents*. 1.

Leroy EM, Kumulungui B, Pourrut X, Rouquet P, Hassanin A, Yaba P, Délicat A, Paweska JT, Gonzalez JP, Swanepoel R. 2005. Fruit bats as reservoirs of Ebola virus. *Nature*. 438:575–576.

Leroy EM, Telfer P, Kumulungui B, Yaba P, Rouquet P, Roques P, Gonzalez JP, Ksiazek TG, Rollin PE, Nerrienet E. 2004. A Serological Survey of Ebola Virus Infection in Central African Nonhuman Primates. *Journal of Infectious Diseases*. 190:1895–1899.

Lewontin RC. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics*. 49:49–67. PMID: 17248194 PMCID: PMC1210557.

Li CX, Shi M, Tian JH, Lin XD, Kang YJ, Chen LJ, Qin XC, Xu J, Holmes EC, Zhang YZ. 2015. Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *eLife*. 4:e05378.

Li KS, Guan Y, Wang J, et al. (22 co-authors). 2004. Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. *Nature*. 430:209–213.

Li W, Shi Z, Yu M, et al. (17 co-authors). 2005. Bats Are Natural Reservoirs of SARS-Like Coronaviruses. *Science*. 310:676–679.

Lindstrom SE, Hiromoto Y, Nishimura H, Saito T, Nerome R, Nerome K. 1999. Comparative analysis of evolutionary mechanisms of the hemagglutinin and three internal protein genes of influenza B virus: Multiple cocirculating lineages and frequent reassortment of the NP, M, and NS genes. *Journal of Virology*. 73:4413–4426.

Longdon B, Hadfield JD, Webster CL, Obbard DJ, Jiggins FM. 2011. Host phylogeny determines viral persistence and replication in novel hosts. *PLoS Pathog*. 7:e1002260.

Lu L, Lycett SJ, Brown AJL. 2014. Reassortment patterns of avian influenza virus internal segments among different subtypes. *BMC Evolutionary Biology*. 14:16.

Lycett SJ, Baillie G, Coulter E, et al. (11 co-authors). 2012. Estimating reassortment rates in co-circulating Eurasian swine influenza viruses. *Journal of General Virology*. 93:2326–2336.

Maganga GD, Kapetshi J, Berthet N, et al. (20 co-authors). 2014. Ebola Virus Disease in the Democratic Republic of Congo. *New England Journal of Medicine*. 371:2083–2091.

Makino S, Keck JG, Stohlman SA, Lai MM. 1986. High-frequency RNA recombination of murine coronaviruses. *Journal of Virology*. 57:729–737.

Margoliash E. 1963. Primary structure and evolution of cytochrome C. *Proceedings of the National Academy of Sciences of the United States of America*. 50:672–679.

Marshall N, Priyamvada L, Ende Z, Steel J, Lowen AC. 2013. Influenza virus reassortment occurs with high frequency in the absence of segment mismatch. *PLoS Pathog*. 9:e1003421.

Maynard Smith J, Smith NH. 1998. Detecting recombination from gene trees. *Molecular Biology and Evolution*. 15:590–599.

McGeoch D, Fellner P, Newton C. 1976. Influenza virus genome consists of eight distinct RNA species. *Proceedings of the National Academy of Sciences of the United States of America*. 73:3045–3049.

McVean G, Awadalla P, Fearnhead P. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*. 160:1231–1241.

Memish ZA, Zumla AI, Al-Hakeem RF, Al-Rabeeah AA, Stephens GM. 2013. Family cluster of Middle East Respiratory Syndrome Coronavirus infections. *New England Journal of Medicine*. 368:2487–2494.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*. 21:1087–1092.

Milkman R, Bridges MM. 1990. Molecular evolution of the Escherichia coli chromosome. III. Clonal frames. *Genetics*. 126:505–517.

Minin VN, Bloomquist EW, Suchard MA. 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution*. 25:1459–1471.

Minin VN, Suchard MA. 2008. Counting labeled transitions in continuous-time Markov models of evolution. *Journal of Mathematical Biology*. 56:391–412.

Mitnaul LJ, Matrosovich MN, Castrucci MR, Tuzikov AB, Bovin NV, Kobasa D, Kawaoka Y. 2000. Balanced hemagglutinin and neuraminidase activities are critical for efficient replication of influenza A virus. *Journal of Virology*. 74:6015–6020.

Miyashita N, Langley CH. 1988. Molecular and phenotypic variation of the white locus region in Drosophila melanogaster. *Genetics*. 120:199–212.

Muller H. 1942. Isolating mechanisms, evolution and temperature. In: Biol. Symp. volume 6, pp. 71–125.

Muller HJ. 1932. Some genetic aspects of sex. *The American Naturalist*. 66:118–138.

Müller MA, Corman VM, Jores J, et al. (12 co-authors). 2014. MERS coronavirus neutralizing antibodies in camels, eastern Africa, 1983–1997. *Emerging Infectious Diseases*. 20.

Müller MA, Meyer B, Corman VM, et al. (19 co-authors). 2015. Presence of Middle East respiratory syndrome coronavirus antibodies in Saudi Arabia: a nationwide, cross-sectional, serological study. *The Lancet Infectious Diseases*. 15:559–564.

Nagarajan N, Kingsford C. 2011. GiRaF: robust, computational identification of influenza reassortments via graph mining. *Nucleic Acids Research*. 39:e34.

Nakagawa N, Nukuzuma S, Haratome S, Go S, Nakagawa T, Hayashi K. 2002. Emergence of an influenza B virus with antigenic change. *Journal of Clinical Microbiology*. 40:3068–3070.

Nath ST, Nayak DP. 1990. Function of two discrete regions is required for nuclear localization of polymerase basic protein 1 of A/WSN/33 influenza virus (H1N1). *Molecular and Cellular Biology*. 10:4139–4145. PMID: 2196448.

Nelson MI, Simonsen L, Viboud C, Miller MA, Holmes EC. 2007. Phylogenetic analysis reveals the global migration of seasonal influenza A viruses. *PLoS Pathog*. 3:e131.

Nerome R, Hiromoto Y, Sugita S, Tanabe N, Ishida M, Matsumoto M, Lindstrom SE, Takahashi T, Nerome K. 1998. Evolutionary characteristics of influenza B virus since its first isolation in 1940: dynamic circulation of deletion and insertion mechanism. *Archives of Virology*. 143:1569–1583.

O'Brien JD, Minin VN, Suchard MA. 2009. Learning to count: Robust estimates for labeled distances between molecular sequences. *Molecular Biology and Evolution*. 26:801–814. PMID: 19131426.

Olabode AS, Jiang X, Robertson DL, Lovell SC. 2015. Ebolavirus is evolving but not changing: No evidence for functional change in EBOV from 1976 to the 2014 outbreak. *Virology*. 482:202–207.

Olal D, Kuehne AI, Bale S, et al. (11 co-authors). 2012. Structure of an Antibody in Complex with Its Mucin Domain Linear Epitope That Is Protective against Ebola Virus. *Journal of Virology*. 86:2809–2816.

Osterhaus ADME, Rimmelzwaan GF, Martina BEE, Bestebroer TM, Fouchier RaM. 2000. Influenza B virus in seals. *Science*. 288:1051–1053.

Parrish CR, Holmes EC, Morens DM, Park EC, Burke DS, Calisher CH, Laughlin CA, Saif LJ, Daszak P. 2008. Cross-Species Virus Transmission and the Emergence of New Epidemic Diseases. *Microbiology and Molecular Biology Reviews*. 72:457–470.

Peiris J, Lai S, Poon L, et al. (16 co-authors). 2003. Coronavirus as a possible cause of severe acute respiratory syndrome. *The Lancet*. 361:1319–1325.

Penny D. 2005. Evolutionary biology: Relativity for molecular clocks. *Nature*. 436:183–184.

Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics*. 21:676–679.

Posada D, Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *Proceedings of the National Academy of Sciences*. 98:13757–13762.

Posada D, Crandall KA. 2002. The Effect of Recombination on the Accuracy of Phylogeny Estimation. *Journal of Molecular Evolution*. 54:396–402.

Posada D, Crandall KA, Holmes EC. 2002. Recombination in Evolutionary Genomics. *Annual Review of Genetics*. 36:75–97.

Pourrut X, Kumulungui B, Wittmann T, Moussavou G, Délicat A, Yaba P, Nkoghe D, Gonzalez JP, Leroy EM. 2005. The natural history of Ebola virus in Africa. *Microbes and Infection*. 7:1005–1014.

Presgraves DC. 2010. The molecular evolutionary basis of species formation. *Nature Reviews Genetics*. 11:175–180.

Presti RM, Zhao G, Beatty WL, Mihindukulasuriya KA, da Rosa APAT, Popov VL, Tesh RB, Virgin HW, Wang D. 2009. Quaranfil, Johnston Atoll, and Lake Chad viruses are novel members of the family Orthomyxoviridae. *Journal of Virology*. 83:11599–11606.

Pupko T, Pe I, Shamir R, Graur D. 2000. A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences. *Molecular Biology and Evolution*. 17:890–896.

Pybus OG, Rambaut A. 2009. Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews Genetics*. 10:540–550.

Pybus OG, Rambaut A, Belshaw R, Freckleton RP, Drummond AJ, Holmes EC. 2007. Phylogenetic Evidence for Deleterious Mutation Load in RNA Viruses and Its Contribution to Viral Evolution. *Molecular Biology and Evolution*. 24:845–852.

Pybus OG, Rambaut A, Harvey PH. 2000. An Integrated Framework for the Inference of Viral Population History From Reconstructed Genealogies. *Genetics*. 155:1429–1437.

Qin Y, Horby PW, Tsang TK, et al. (25 co-authors). 2015. Differences in the epidemiology of human cases of avian influenza A(H7N9) and A(H5N1) viruses infection. *Clinical Infectious Diseases*. p. civ345.

Rambaut A, Holmes E. 2009. The early molecular epidemiology of the swine-origin A/H1N1 human influenza pandemic. *PLoS Currents*. 1.

Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC. 2008. The genomic and epidemiological dynamics of human influenza A virus. *Nature*. 453:615–619.

Rambaut A, Suchard M, Drummond A. 2009. Tracer v1.5. Available at http://tree.bio.ed.ac.uk/software/tracer/.

Ran Z, Shen H, Lang Y, et al. (19 co-authors). 2015. Domestic Pigs Are Susceptible to Infection with Influenza B Viruses. *Journal of Virology*. 89:4818–4826.

Rannala B, Yang Z. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution*. 43:304–311.

Ratner V, Kolchanov N, Omel'ianchuk L. 1989. [Phylogenetic analysis of genes of the influenza virus. Relationship between adaptability and neutrality]. *Genetika*. 25:1499–1507.

Reed C, Meltzer MI, Finelli L, Fiore A. 2012. Public health impact of including two lineages of influenza B in a quadrivalent seasonal influenza vaccine. *Vaccine*. 30:1993–1998.

Reusken CB, Haagmans BL, Müller MA, et al. (24 co-authors). 2013. Middle East Respiratory Syndrome Coronavirus neutralising serum antibodies in dromedary camels: a comparative serological study. *The Lancet Infectious Diseases*. 13:859–866.

Reusken CB, Messadi L, Feyisa A, et al. (17 co-authors). 2014. Geographic distribution of MERS coronavirus among dromedary camels, Africa. *Emerging Infectious Diseases*. 20:1370–1374.

Robertson DL, Hahn BH, Sharp PM. 1995a. Recombination in AIDS viruses. *Journal of Molecular Evolution*. 40:249–259.

Robertson DL, Sharp PM, McCutchan FE, Hahn BH. 1995b. Recombination in HIV-1. *Nature*. 374:124–126.

Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, Feil EJ. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. *Journal of Theoretical Biology*. 239:226–235.

Rodriguez LL, Roo AD, Guimard Y, et al. (13 co-authors). 1999. Persistence and Genetic Stability of Ebola Virus during the Outbreak in Kikwit, Democratic Republic of the Congo, 1995. *Journal of Infectious Diseases*. 179:S170–S176.

Rota JS, Hummel KB, Rota PA, Bellini WJ. 1992. Genetic variability of the glycoprotein genes of current wild-type measles isolates. *Virology*. 188:135–142.

Rota PA, Wallis TR, Harmon MW, Rota JS, Kendal AP, Nerome K. 1990. Cocirculation of two distinct evolutionary lineages of influenza type B virus since 1983. *Virology*. 175:59–68.

Ruane S, Pyron RA, Burbrink FT. 2011. Phylogenetic relationships of the Cretaceous frog Beelzebufo from Madagascar and the placement of fossil constraints based on temporal and phylogenetic evidence. *Journal of Evolutionary Biology*. 24:274–285.

Ruigrok RW, Crépin T, Kolakofsky D. 2011. Nucleoproteins and nucleocapsids of negative-strand RNA viruses. *Current Opinion in Microbiology*. 14:504–510.

Russell CA, Jones TC, Barr IG, et al. (28 co-authors). 2008. The global circulation of seasonal influenza a (H3N2) viruses. *Science*. 320:340–346.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*. 4:406–425. PMID: 3447015.

Sanchez A, Trappier SG, Mahy BW, Peters CJ, Nichol ST. 1996. The virion glycoproteins of Ebola viruses are encoded in two reading frames and are expressed through transcriptional editing. *Proceedings of the National Academy of Sciences*. 93:3602–3607.

Sanderson MJ. 1997. A Nonparametric Approach to Estimating Divergence Times in the Absence of Rate Constancy. *Molecular Biology and Evolution*. 14:1218.

Schierup MH, Hein J. 2000. Consequences of Recombination on Traditional Phylogenetic Analysis. *Genetics*. 156:879–891.

Schoepp RJ, Rossi CA, Khan SH, Goba A, Fair JN. 2014. Undiagnosed Acute Viral Febrile Illnesses, Sierra Leone. *Emerging Infectious Diseases*. 20:1176–1182.

Scholtissek C, Rohde W, Von Hoyningen V, Rott R. 1978a. On the origin of the human influenza virus subtypes H2N2 and H3N2. *Virology*. 87:13–20.

Scholtissek C, von Hoyningen V, Rott R. 1978b. Genetic relatedness between the new 1977 epidemic strains (H1N1) of influenza and human influenza strains isolated between 1947 and 1957 (H1N1). *Virology*. 89:613–617.

Sharp PM, Bailes E, Chaudhuri RR, Rodenburg CM, Santiago MO, Hahn BH. 2001. The origins of acquired immune deficiency syndrome viruses: where and when? *Philosophical Transactions of the Royal Society of London. Series B.* 356:867–876.

Shaw MW, Xu X, Li Y, Normand S, Ueki RT, Kunimoto GY, Hall H, Klimov A, Cox NJ, Subbarao K. 2002. Reappearance and global spread of variants of influenza B/Victoria/2/87 lineage viruses in the 2000–2001 and 2001–2002 seasons. *Virology.* 303:1–8.

Shriner D, Nickle DC, Jensen MA, Mullins JI. 2003. Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genetics Research.* 81:115–121.

Simmonds P. 2012. SSE: a nucleotide and amino acid sequence analysis platform. *BMC research notes.* 5:50.

Slatkin M, Maddison WP. 1989. A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics.* 123:603–613.

Smith GJD, Vijaykrishna D, Bahl J, et al. (13 co-authors). 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature.* 459:1122–1125.

Smith JS, Orciari LA, Yager PA. 1995. Molecular epidemiology of rabies in the United States. *Seminars in Virology.* 6:387–400.

Smith SA, Donoghue MJ. 2008. Rates of Molecular Evolution Are Linked to Life History in Flowering Plants. *Science.* 322:86–89.

Squires RB, Noronha J, Hunt V, et al. (17 co-authors). 2012. Influenza Research Database: an integrated bioinformatics resource for influenza research and surveillance. *Influenza and Other Respiratory Viruses.* 6:404–416.

Steinhauer DA, Domingo E, Holland JJ. 1992. Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase. *Gene.* 122:281–288.

Subbarao EK, London W, Murphy BR. 1993. A single amino acid in the PB2 gene of influenza A virus is a determinant of host range. *Journal of Virology.* 67:1761–1764.

Svinti V, Cotton JA, McInerney JO. 2013. New approaches for unravelling reassortment pathways. *BMC Evolutionary Biology*. 13:1. PMID: 23279962.

Swanepoel R, Leman PA, Burt FJ, Zachariades NA, Braack LE, Ksiazek TG, Rollin PE, Zaki SR, Peters CJ. 1996. Experimental inoculation of plants and animals with Ebola virus. *Emerging Infectious Diseases*. 2:321–325.

Tavaré S. 1986. Some Mathematical Questions in Biology: DNA Sequence Analysis. Lectures on Mathematics in the Life Sciences, volume 17, 57-86. American Mathematical Society.

Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution*. 15:1647–1657.

Tisdale M, Kemp SD, Parry NR, Larder BA. 1993. Rapid in vitro selection of human immunodeficiency virus type 1 resistant to 3'-thiacytidine inhibitors due to a mutation in the YMDD region of reverse transcriptase. *Proceedings of the National Academy of Sciences of the United States of America*. 90:5653–5656.

Tong YG, Shi WF, Liu D, et al. (56 co-authors). 2015. Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. *Nature*. 524:93–96.

Turner TL, Hahn MW, Nuzhdin SV. 2005. Genomic islands of speciation in Anopheles gambiae. *PLoS Biol*. 3:e285.

Uzicanin A, Zimmerman L. 2011. Field Effectiveness of Live Attenuated Measles-Containing Vaccines: A Review of Published Literature. *Journal of Infectious Diseases*. 204:S133–S149.

Vijaykrishna D, Holmes EC, Joseph U, et al. (28 co-authors). 2015. The contrasting phylodynamics of human influenza B viruses. *eLife*. 4:e05055.

Villarreal LP, Defilippis VR, Gottlieb KA. 2000. Acute and Persistent Viral Life Strategies and Their Relationship to Emerging Diseases. *Virology*. 272:1–6.

Volchkov VE, Becker S, Volchkova VA, Ternovoj VA, Kotov AN, Netesov SV, Klenk HD. 1995. GP mRNA of Ebola Virus Is Edited by the Ebola Virus Polymerase and by T7 and Vaccinia Virus Polymerases. *Virology*. 214:421–430.

Ward MJ, Lycett SJ, Kalish ML, Rambaut A, Brown AJL. 2013. Estimating the rate of intersubtype recombination in early HIV-1 group M strains. *Journal of Virology.* 87:1967–1973.

Weinert LA, Welch JJ, Suchard MA, Lemey P, Rambaut A, Fitzgerald JR. 2012. Molecular dating of human-to-bovid host jumps by Staphylococcus aureus reveals an association with the spread of domestication. *Biology Letters.* 8:829–832.

Whidden C, Beiko RG, Zeh N. 2010. Fast FPT algorithms for computing rooted agreement forests: Theory and experiments. In: Festa P, editor, Experimental Algorithms, Springer Berlin Heidelberg, number 6049 in Lecture Notes in Computer Science, pp. 141–153.

Whidden C, Beiko RG, Zeh N. 2013. Fixed-parameter algorithms for maximum agreement forests. *SIAM Journal on Computing.* 42:1431–1466.

Whidden C, Zeh N. 2009. A unifying view on approximation and FPT of agreement forests. In: Salzberg SL, Warnow T, editors, Algorithms in Bioinformatics, Springer Berlin Heidelberg, number 5724 in Lecture Notes in Computer Science, pp. 390–402.

Woolhouse MEJ, Haydon DT, Antia R. 2005. Emerging pathogens: the epidemiology and evolution of species jumps. *Trends in Ecology & Evolution.* 20:238–244.

World Health Organization. 2009. Influenza Fact sheet. Available at http://www.who.int/mediacentre/factsheets/fs211/en/.

Worobey M, Han GZ, Rambaut A. 2014. A synchronized global sweep of the internal genes of modern avian influenza virus. *Nature.* advance online publication.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution.* 39:306–314.

Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. 2012. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *New England Journal of Medicine.* 367:1814–1820.

Zhao H, Nettleton D, Soller M, Dekkers JCM. 2005. Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. *Genetics Research*. 86:77–87.

Zuckerkandl E, Pauling L. 1962. Molecular disease, evolution and genetic heterogeneity. In: Horizons in Biochemistry, Academic Press, pp. 189–225.

Appendices

APPENDIX A

MERS-COV RECOMBINATION: IMPLICATIONS ABOUT THE RESERVOIR AND POTENTIAL FOR ADAPTATION

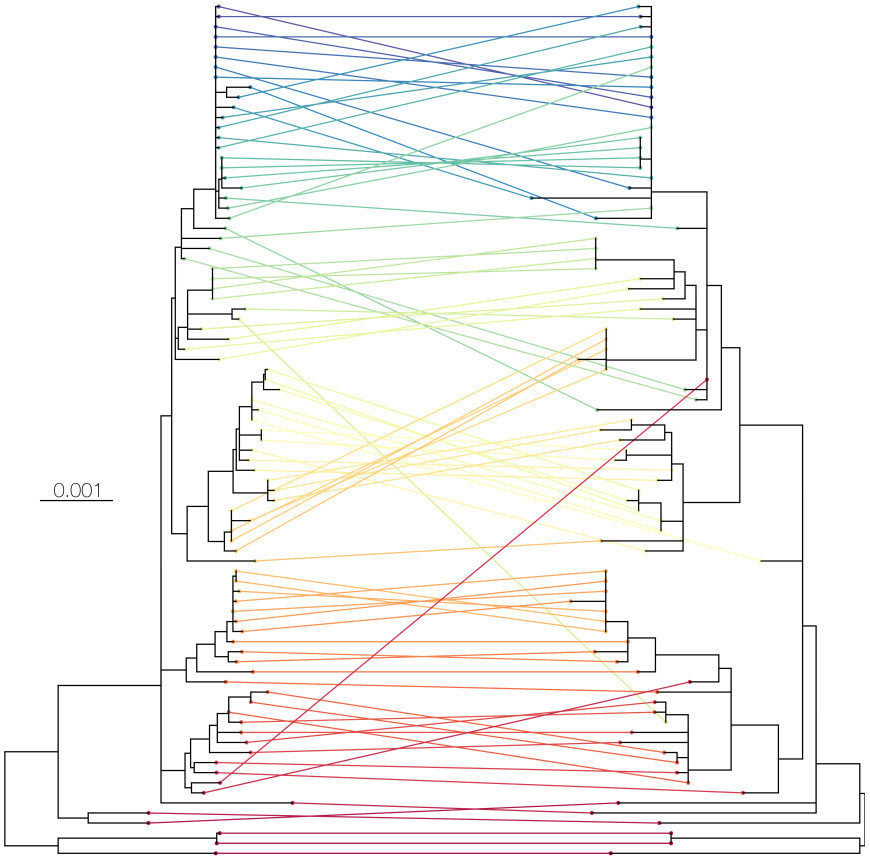


FIGURE A.1: **Trees recovered using GARD (Kosakovsky Pond et al., 2006) across the breakpoint at position 23722.** NJ trees reconstructed by GARD across the first identified breakpoint. Tree from positions 1-23722 on the left and positions 23723-30126 on the right. The same tips in both trees are connected by coloured lines to indicate phylogenetic incongruity.

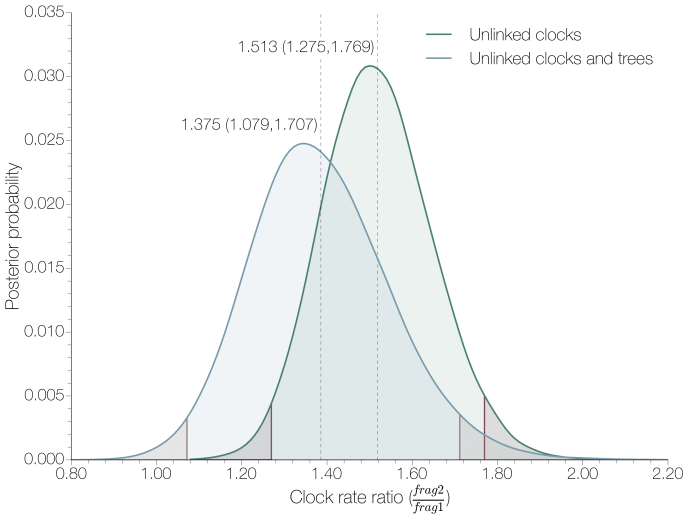


FIGURE A.2: **Empirical rate heterogeneity in MERS-CoV genome.** Posterior estimates of the ratio between the molecular clock rates estimated independently from GARD-inferred fragment 2 (positions 23723-30126) and fragment 1 (positions 1-23722) under independent or linked tree models derived from 3 independent marginal likelihood analyses. Dotted lines indicate the mean of the distribution and numbers next to the line show the median and the 95% highest posterior density intervals.

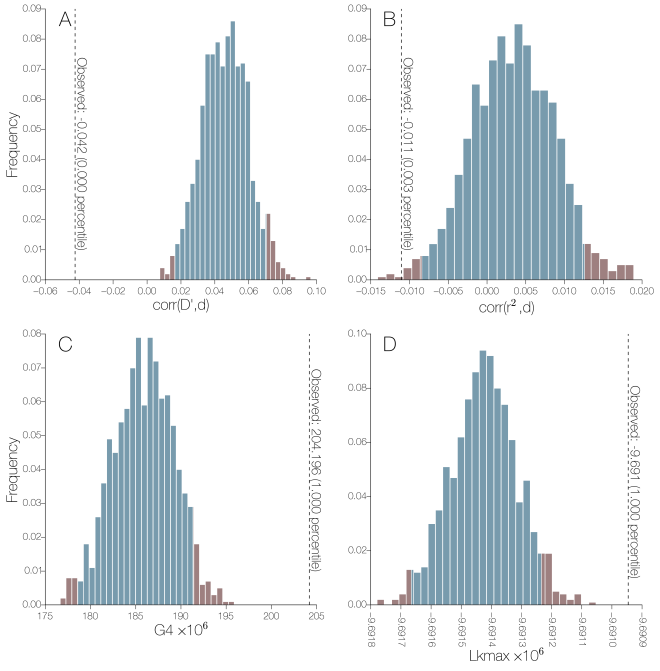


FIGURE A.3: **LDhat (McVean et al., 2002) permutation test results for MERS-CoV.** All 4 observed LD decay statistics (A - corr(D',d), B - corr(r^2,d), C - G4, D - Lkmax) for MERS-CoV data fall outside the distribution generated by permuting sites in ways consistent with recombination.

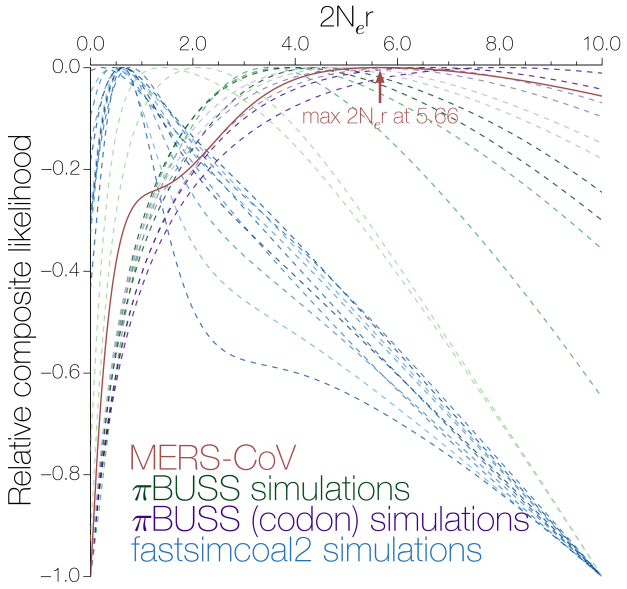


FIGURE A.4: **Relative composite likelihood surface.** Composite likelihoods for the recombination rate estimates were rescaled to be within the range [-1,0]. Surfaces are coloured by data source: MERS-CoV estimate is in red, πBUSS simulations under a nucleotide substitution model in green, πBUSS simulations under a codon substitution model in purple and fastsimcoal2 simulations in blue. Colour scheme is identical to figure 3.2 in the main text. Maximum composite likelihood for MERS-CoV data is achieved at ρ=5.66, all other datasets have an inferred recombination rate above 0 despite being simulated without recombination.

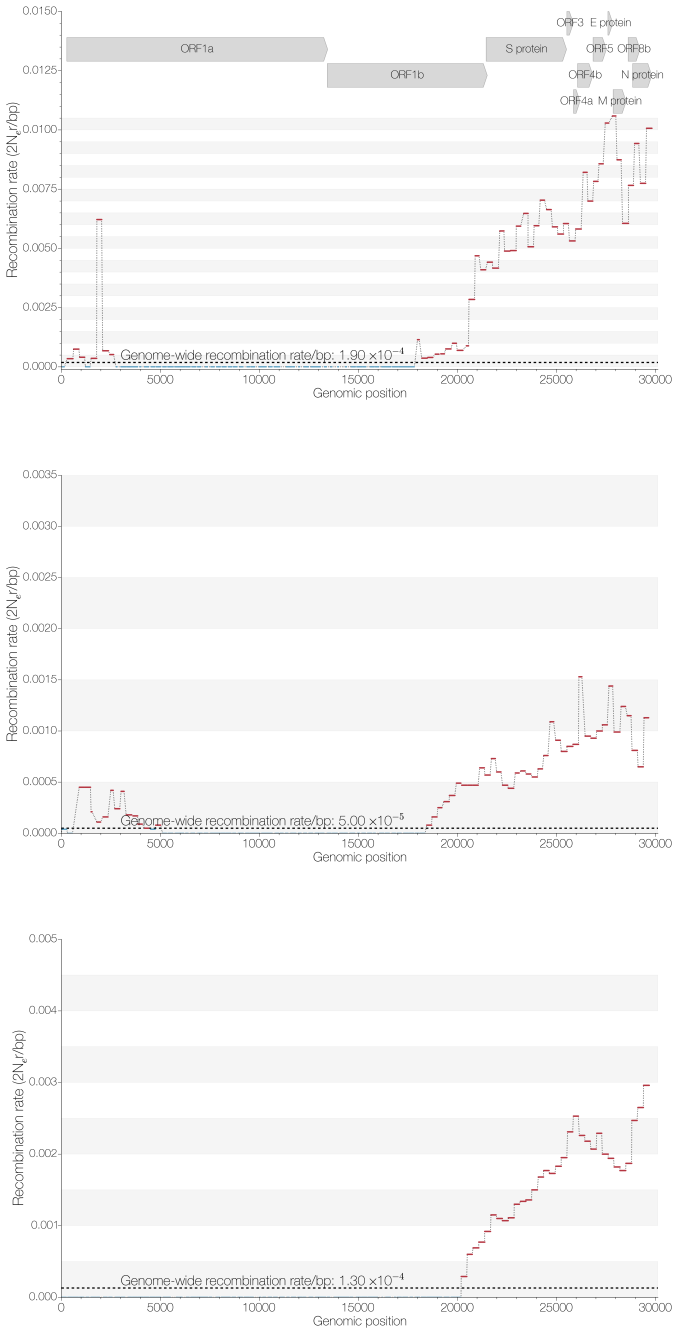


FIGURE A.5: **Window-based estimates of recombination rate.** Inferred recombination rates for 300 nucleotide-long windows in MERS-CoV genome (top), πBUSS-simulated sequences with 1.3× rate heterogeneity (middle) and 3× rate heterogeneity (bottom) under a nucleotide substitution model. Recombination rates that are above the inferred genome-wide recombination rate are in red. Simulated rate heterogeneity is sufficient to mislead this method, although the inferred recombination rates in the last third of the MERS-CoV genome are much greater than those inferred from the simulated data.

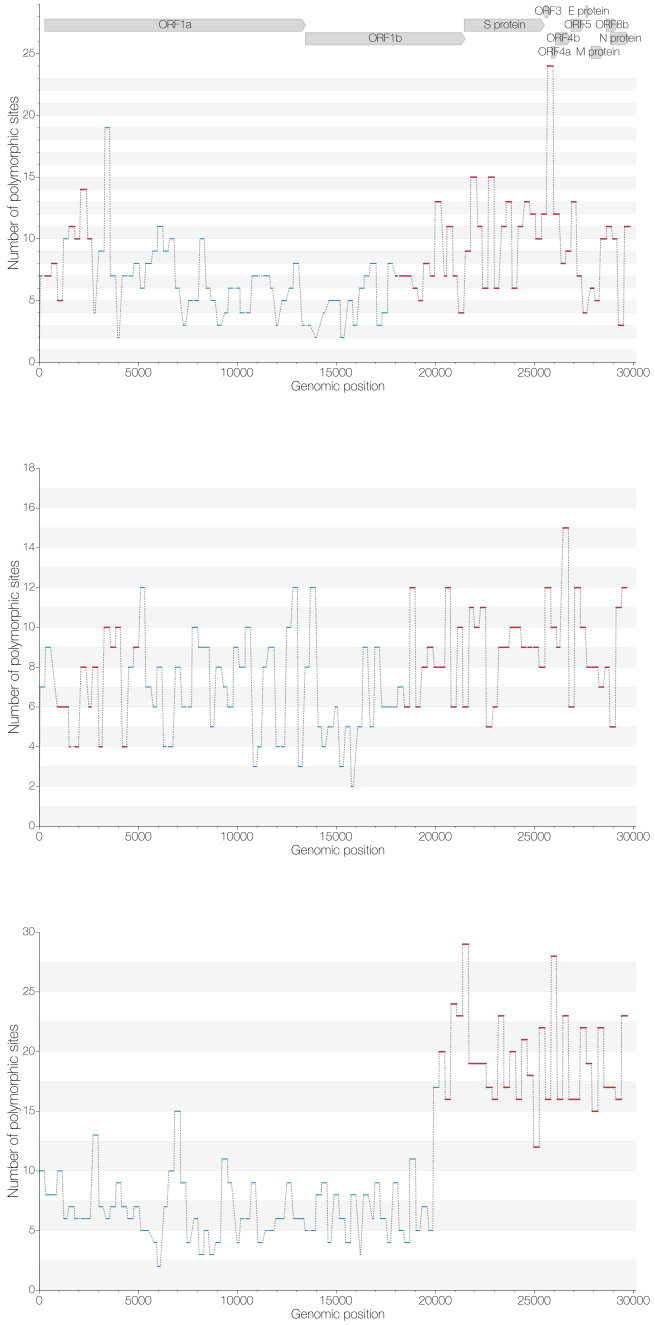


FIGURE A.6: **Window-based estimates of polymorphic site density.** Inferred polymorphic site densities for 300 nucleotide-long windows in MERS-CoV genome (top), πBUSS-simulated sequences with 1.3× rate heterogeneity (middle) and 3× rate heterogeneity (bottom) under a nucleotide substitution model. Windows are coloured red if their recombination rate is above the inferred genome-wide recombination rate. Extreme rate heterogeneity (3×) results in a higher density of polymorphic sites in the region with the higher rate.

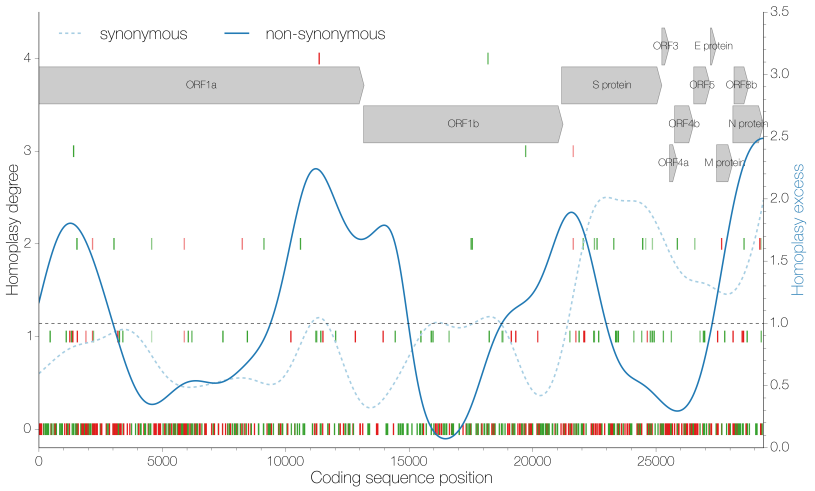


FIGURE A.7: **Homoplasy degrees inferred by BEAST (Drummond et al., 2012).** Position along the genome is shown on the x axis and homoplasy degree, the number of times a particular mutation has occured in excess in the tree, is shown on the y axis. Individual mutations are marked by vertical lines, synonymous ones in green and non-synonymous in red with transparency representing the posterior probability of a given homoplasy degree for each mutation. The ratio of apparent homoplasy over synapomorphy kernel density estimates (bandwidth=0.1) is shown in blue for synonymous (dashed) and non-synonymous (solid) sites separately. Arrows at the top indicate the positions and names of coding sequences within the MERS-CoV genome.

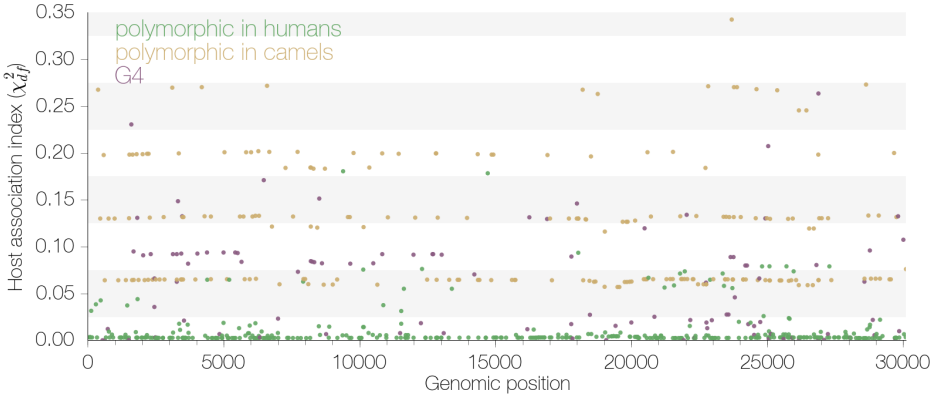


FIGURE A.8: **Host association indices for variable sites.** Estimates for the association between particular alleles and host. The association index is an adapted version of the χ^2_{df} statistic of LD (Hedrick and Thomson, 1986), and quantifies how well one can predict the allele at any given polymorphic site, given the host it was isolated from. No perfect associations (association index = 1.0) between particular alleles and host (human or camel) were found.

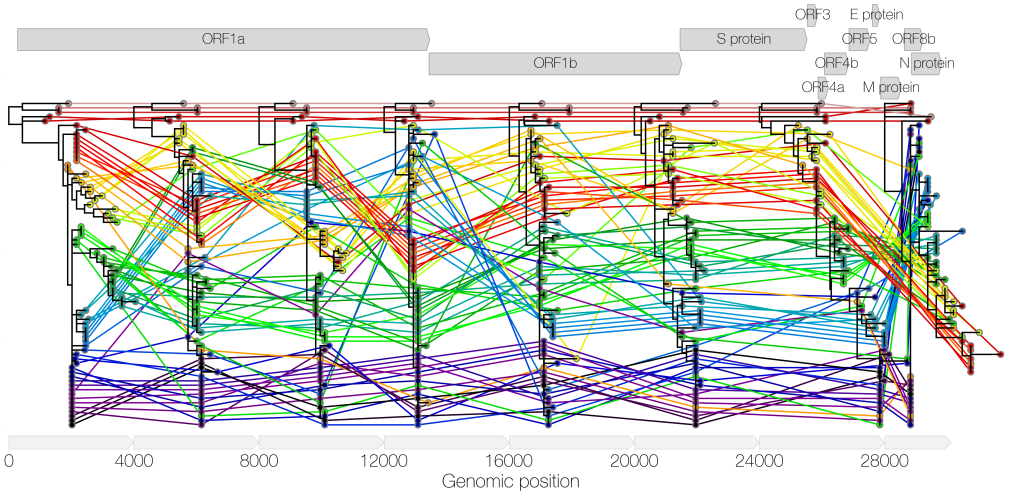


FIGURE A.9: **Maximum likelihood phylogenies across MERS-CoV genome.** Maximum likelihood phylogenies recovered with PhyML (Guindon and Gascuel, 2003) under GTR+Γ_4 (Tavaré, 1986; Yang, 1994) nucleotide substitution model across 4000 nucleotide fragments derived from the MERS-CoV genome. Each tip is connected to its counterpart in phylogenies of neighboring fragments and coloured sequentially according to the order in which tips appear in the first fragment. Arrows at the top indicate the relative positions, lengths and names of coding sequences in the MERS-CoV genome, arrows at the bottom indicate the relative lengths of fragments used to produce the phylogenies.

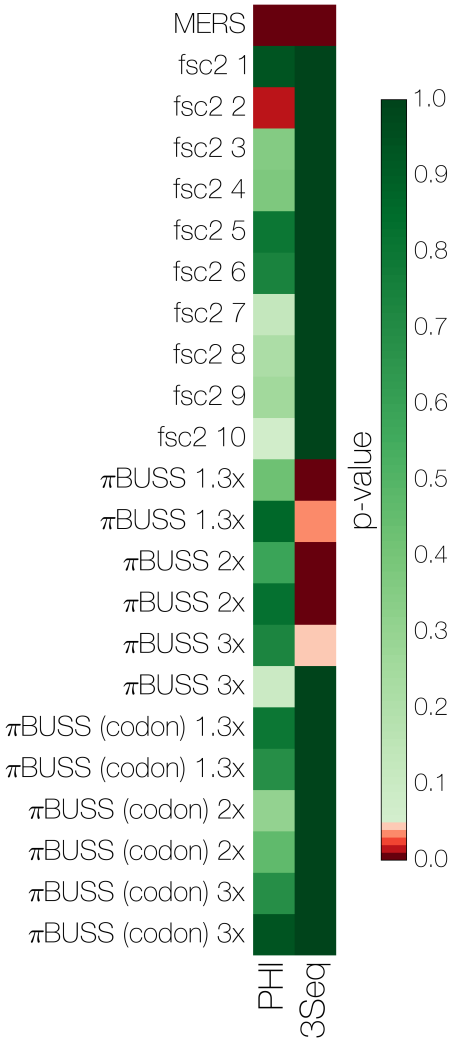


FIGURE A.10: **Results (p-values) from pairwise homoplasy index (PHI) (Bruen et al., 2006) and 3Seq (Boni et al., 2007) analyses on MERS and simulated datasets.** Both PHI and 3Seq analyses indicate that there is strong evidence of recombination in MERS-CoV (PHI p-value and Bonferroni-corrected 3Seq p-values <0.05, in red). Some simulated datasets are spuriously identified as recombinant by either PHI or 3Seq, but not both.

The following is a list of MERS-CoV strains used in the study with GenBank accessions indicated where available.

Camel-Egypt-NRCE-HKU270—Egypt—2013-12-30	
Camel-Egypt-NRCE-HKU271—Egypt—2013-12-30	
Camel-KFU-HKU-13—KSA—unknown—KJ650295—2013-12-30	KJ650295
Camel-KFU-HKU-19Dam—KSA—unknown—KJ650296—2013-12-30	KJ650296
Camel-KFU-HKU-1—KSA—unknown—KJ650297—2013-11-30	KJ650297
Al-Hasa-17-2013—Dammam—2013-05-15	KF600647
Al-Hasa-18-2013—Dammam—2013-05-23	KF600651
Al-Hasa-19-2013—Dammam—2013-05-23	KF600632
Al-Hasa-21-2013—Al-Hofuf—2013-05-30	KF600634
Al-Hasa-24-2013—Al-Hofuf—2013-05-01	KJ156867, KJ156870, KJ156875, KJ156885, KJ156902, KJ156919
Al-Hasa-25-2013—Al-Hofuf—2013-05-02	KJ156866
Al-Hasa-26-2013—Dhahran—2013-06-18	KJ156872, KJ156882, KJ156941
Al-Hasa-27-2013—Al-Hasa—2013-06-19	KJ156943, KJ156939
Al-Hasa-28-2013—Al-Hofuf—2013-06-22	KJ156884, KJ156887, KJ156889, KJ156893, KJ156909, KJ156930, KJ156940
Asir-1-2013—Khamis-Mushait—2013-07-02	KJ156948, KJ156883, KJ156903, KJ156925, KJ156948

Asir-2-2013—Asir—2013-08-05	KJ156863, KJ156898, KJ156899, KJ156900, KJ156912, KJ156932, KJ156945
B-Al-Hasa-12-2013—Al-Hofuf—2013-05-07	KF600627
Buraidah-1-2013—Buraidah—2013-05-13	KF600630
Camel-Al-Hasa-KFU-1-2013—KSA—2013-11-30	KJ650295
Camel-Egypt-NRCE-HKU205—Nile—2013-11-06	KJ477102
Camel-KSA-503—KJ713297—Taif—2013-11	KJ713297
Camel-KSA-505—KJ713295—Taif—2013-11	KJ713295
Camel-KSA-363—KJ713298—Taif—2013-11	KJ713298
Camel-KSA-376—KJ713299—Taif—2013-11	KJ713299
Camel-KSA-378—KJ713296—Taif—2013-11	KJ713296
Camel-Oman30-2013—Oman—2013-12	KJ573789, KJ573790, KJ573791, KJ573792, KJ573793
Camel-Qatar-1-2013—Qatar—2013-10-17	KF933385
Camel-Qatar-2—KJ650098—Qatar—2014-02-16	KJ650098
EMC-2012—Bisha—2012-06-13	JX869059
England-Qatar-2012—Doha—2012-09-19	KC667074, KC164505
England2-HPA-2013—Makkah—2013-02-10	KM015348
F-Al-Hasa-9-2013—Al-Hofuf—2013-05-01	KF600622, KF600639, KF600648, KF600649, KF600654

Florida/USA-2-Saudi-Arabia—KSA—Jeddah—2014-05-10	KJ829365
France-1-2013—KF745068—UAE—2013-05-07	KF745068
G-Al-Hasa-11-2013—Al-Hofuf—2013-05-03	KF600629, KF600636, KF600646
Greece-1-Saudi-Arabia—KSA—2014-04-18	KJ782549, KJ782550
H-Al-Hasa-7-2013—Al-Hofuf—2013-05-01	KF600623, KF600655
Hafr-Al-Batin-1-2013—Hafr-Al-Batin—2013-06-04	KF600628
Hafr-Al-Batin-2-2013—Hafr-Al-Batin—2013-08-05	KJ156910
Hafr-Al-Batin-4-2013—Hafr-Al-Batin—2013-08-25	KJ156861, KJ156864, KJ156895, KJ156931
Hafr-Al-Batin-5-2013—Hafr-Al-Batin—2013-08-25	KJ156913, KJ156924, KJ156951, KJ156954
Hafr-Al-Batin-6-2013—Hafr-Al-Batin—2013-08-28	KJ156874
I-Al-Hasa-3-2013—Al-Hofuf—2013-04-22	KF600613
Indiana/USA-1-Saudi-Arabia—KSA—Riyadh—KJ813439—2014-04-30	KJ813439
J-Al-Hasa-2-2013—Al-Hofuf—2013-04-21	KF186566
Jeddah-1-2013—KSA—2013-11-06	KJ556336
Jordan—3-2012—Al-Zarqa—2012-04-15	KC776174
K-Al-Hasa-4-2013—Al-Hofuf—2013-05-01	KF186564

M-Al-Hasa-10-2013—Al-Hofuf—2013-05-02	KF600614, KF600624, KF600641, KF600642, KF600653
Madinah-1-2013—Medina—2013-09-01	KJ156935, KJ156904, KJ156917
Madinah-3-2013—Medina—2013-09-11	KJ156950, KJ156916
Munich-2013—Abu-Dhabi—2013-03-22	KF192507
Netherlands-1-2014—KSA——2014	KJ858497, KJ858498, KJ858499, KJ858500
Netherlands-2-2014—KSA——2014	KJ858495, KJ858496
O-Al-Hasa-8-2013—Al-Hofuf—2013-05-01	KF600618, KF600626, KF600635, KF600638
Qatar-3-2013—KF961221—Qatar—2013-10-01	KF961221
Qatar-4-2013—KF961222—Qatar—2013-10-01	KF961222
Riyadh-10-2013—Riyadh—2013-08-05	KJ156891, KJ156907, KJ156936
Riyadh-11-2013—Riyadh—2013-08-06	KJ156946, KJ156911
Riyadh-12-2013—Riyadh—2013-08-08	KJ156926, KJ156901
Riyadh-13-2013—Riyadh—2013-08-13	KJ156873, KJ156888
Riyadh-14-2013—Riyadh—2013-08-15	KJ156934
Riyadh-15-2013—Riyadh—2013-08-19	KJ156859, KJ156877, KJ156878, KJ156914
Riyadh-17-2013—Riyadh—2013-08-26	KJ156918, KJ156920, KJ156865
Riyadh-1-2012—Riyadh—2012-10-23	KF600612
Riyadh-2-2012—Riyadh—2012-10-30	KF600652
Riyadh-3-2013—Riyadh—2013-02-05	KF600613
Riyadh-4-2013—Riyadh—2013-05-01	KJ156952
Riyadh-5-2013—Riyadh—2013-07-02	KJ156944

Riyadh-6-2013—Riyadh—2013-07-02	KJ156947, KJ156879, KJ156890, KJ156908
Riyadh-7-2013—Riyadh—2013-07-15	KJ156937, KJ156905
Riyadh-8-2013—Riyadh—2013-07-17	KJ156942, KJ156880
Riyadh-9-2013—Riyadh—2013-07-17	KJ156869
S-Al-Hasa-14-2013—Al-Hofuf—2013-05-08	KF600615, KF600643
T-Al-Hasa-16-2013—Al-Hofuf—2013-05-12	KF600644
Taif-1-2013—Taif—2013-06-12	KJ156949
Taif-2-2013—Taif—2013-06-13	KJ156876, KJ156896
Taif-3-2013—Taif—2013-06-13	KJ156897, KJ156906, KJ156921, KJ156922, KJ156938, KJ156868
Taif-4-2013—Taif—2013-06-13	KJ156871, KJ156886
Tunisia-Qatar-2013—Qatar—2013-05-08	KF811035, KF811036
V-Al-Hasa-1-2013—Dammam—2013-05-09	KF186567
W-Al-Hasa-23-2013—Al-Hofuf—2013-05-13	KJ156860, KJ156862, KJ156894, KJ156929
Wadi-Ad-Dawasir-1-2013—Wadi-Ad-Dawasir—2013-06-12	KJ156881
X-Al-Hasa-22-2013—Dammam—2013-05-09	KF600617, KF600619, KF600621, KF600625, KF600631, KF600633
Y-Al-Hasa-15-2013—Dammam—2013-05-11	KF600645
Jeddah-C8826—KSA—Jeddah—2014-04-12	KM027258
Jeddah-C9055—KSA—Jeddah—2014-04-14	KM027259
Jeddah-C7149—KSA—Jeddah—2014-04-05	KM027255
Jeddah-C7569—KSA—Jeddah—2014-04-03	KM027256
Jeddah-C7770—KSA—Jeddah—2014-04-07	KM027257

Makkah-C9355—KSA—Makkah—2014-04-15	KM027261
Jeddah-C10306—KSA—Jeddah—2014	KM027260
Riyadh-683—KSA—Riyadh—2014	KM027262
Riyadh-693—KSA—Riyadh—2014	KM027291
Riyadh-158—KSA—Riyadh—2014	KM027281
Riyadh-042—KSA—Riyadh—2014	KM027277
Riyadh-057—KSA—Riyadh—2014	KM027278
Riyadh-059—KSA—Riyadh—2014	KM027279
Riyadh-150—KSA—Riyadh—2014	KM027280
Riyadh-166—KSA—Riyadh—2014	KM027282
Riyadh-167—KSA—Riyadh—2014	KM027283
Riyadh-345—KSA—Riyadh—2014	KM027284
Riyadh-349—KSA—Riyadh—2014	KM027285
Riyadh-351—KSA—Riyadh—2014	KM027286
Riyadh-686—KSA—Riyadh—2014	KM027290
Riyadh-359—KSA—Riyadh—2014	KM027287
Riyadh-551—KSA—Riyadh—2014	KM027288
Riyadh-564—KSA—Riyadh—2014	KM027289
Riyadh-695—KSA—Riyadh—2014	KM027292

APPENDIX B

REASSORTMENT BETWEEN INFLUENZA B LINEAGES AND THE EMERGENCE OF A CO-ADAPTED PB1-PB2-HA GENE COMPLEX

Confirmation of primary findings

We sought to confirm our findings through measurement of linkage disequilibrium (LD), a measure of non-random association between polymorphic loci within a population. We estimated LD directly from haplotype frequencies at polymorphic amino acid sites (see Methods) in the secondary dataset, thereby avoiding phylogenetic reconstruction or Vic/Yam lineage assignment. We observe greater amino acid LD values between PB1, PB2 and HA than between other pairs of segments (Figure B.1) in a large secondary dataset. This suggests that PB1, PB2 and HA segments possess a considerable number of co-assorting non-synonymous alleles, which upon closer inspection are associated with either Vic or Yam lineage segments. We conclude that Victoria and Yamagata lineages of PB1, PB2 and HA have accumulated lineage-specific amino acid substitutions. Of the amino acid sites that exhibit high LD on PB1, PB2 and HA proteins, there are 4 sites on PB1, 4 on PB2 and 4 on HA proteins which form a network of sites exhibiting high LD (Figures B.3 and B.2). These sites define the split between Vic and Yam lineages within PB1, PB2 and HA segments. In addition, there are sites on PB1, PB2, HA and NA proteins which also show high, albeit smaller, LD which correspond to sites which have undergone amino acid replacements some time after the Vic/Yam split.

Analysis of within-lineage reassortment patterns

Subtree prune and regraft (SPR) distances between phylogenetic trees are an approximate measure of the numbers of reassortment or recombination events (Svinti et al., 2013). Exact SPR distances are difficult to compute, as they depend on the SPR distance itself and are impractical to compute for posterior distributions of trees except for the most similar trees. We calculated approximate SPR distances (Whidden and Zeh, 2009; Whidden et al., 2010, 2013) to quantify the numbers of reassortments that have taken place between all pairs of segments. Normalized approximate SPR distances,d_{SPR}, were recovered using (see Methods):

$$d_{\mathrm{SPR}}(A_i, B_i) = \frac{f(A_i, A'_i) + f(B_i, B'_i)}{2\, f(A_i, B_i)}, \tag{B.1}$$

where $f(A_i, A'_i)$, $f(B_i, B'_i)$ and $f(A_i, B_i)$ are approximate SPR distances between ith posterior samples from segments A, B and independent analyses thereof (A' and B').

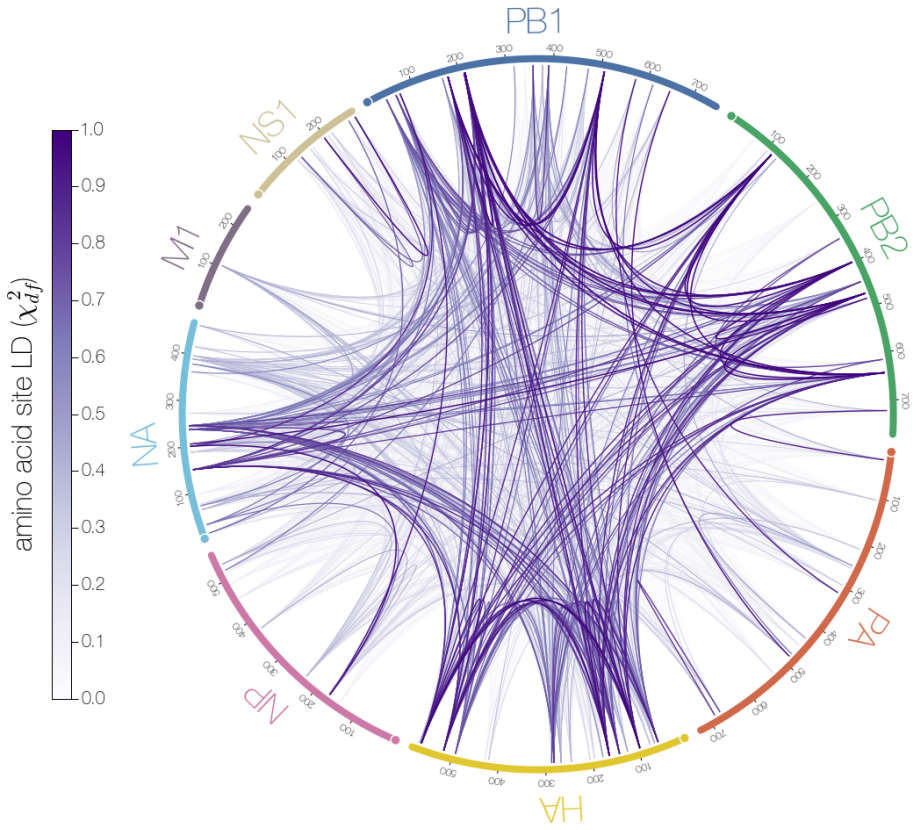


FIGURE B.1: **LD comparison between influenza B proteins.** Pairwise comparisons of linkage disequilibrium between amino acid sites on influenza B proteins in the secondary dataset. Many polymorphic amino acid sites on PB1, PB2 and HA proteins exhibit high LD between themselves, followed by the NA protein. This is evidence of a considerable number of co-assorting alleles within these proteins.

Figure B.4 shows approximate SPR distances between all pairs of segment trees after normalization. If there are biases in the way segments reassort, so that some segments tend to co-assort more often, we expect to observe a lower reassortment rate between them, which would manifest as small-scale similarities between phylogenetic trees of those segments. In our case we expect SPR distances, which are proportional to the number of reassortment events that have taken place between trees, to reflect the overall (*i.e* both within and between lineages) reassortment rate.

The 95% highest posterior density (HPD) intervals of normalized approximate SPR distances between pairs of segments encompass most means and occupy a relatively small range, suggesting there is no evidence of differences in the number of reassortments between segments (Figure B.4). Reassortment rate given as number of SPR moves per total time in both trees shows similar results (Figure B.5). This is in line with recent experiments in influenza A that have shown that reassortment between

FIGURE B.2: **Heatmap of linkage disequilibrium (χ^2_{df}) between amino acid sites on PB1, PB2 and HA proteins.** Numbers next to each row and column correspond to amino acid site number within a given protein starting from methionine. Amino acid sites exhibiting reciprocally high LD between PB1, PB2 and HA proteins are: 176, 211, 213, 508 (PB1), 115, 399, 470, 642 (PB2) and 96, 163, 520 and 570 (HA). Sites 211 and 213 on the PB1 protein are very close to each other and the stretch of sequence around these residues contains many positively charged amino acids (lysine and arginine). Multiple nuclear localization signals (NLSs) are predicted to occur around this region and sites 211 and 213 are either predicted to be near the end of a mono-partite NLS or the beginning of a bi-partite NLS. Previous research (Nath and Nayak, 1990) suggests that in the influenza A PB1 protein residue 211 (homologous to influenza B PB1 residue 211) is the last residue of a bi-partite NLS. Almost all Yamagata lineage isolates possess arginine (R) residue at PB1 position 211 and a serine (S) residue at position 213, whereas Victoria lineage isolates have lysine (K) at position 211 and threonine (T) at position 213. It remains to be seen whether these sites significantly affect the nuclear import efficiency of the PB1 protein of either lineage. Though the PB1 protein is known to accumulate in the nucleus on its own, efficient import into the nucleus requires the presence of the PA protein (Fodor and Smith, 2004). Similarly, site 399 on the PB2 protein are close to residues 377, 406 and 408 which are homologous to sites in influenza A that are responsible for mRNA cap-binding (Guilligay et al., 2008).

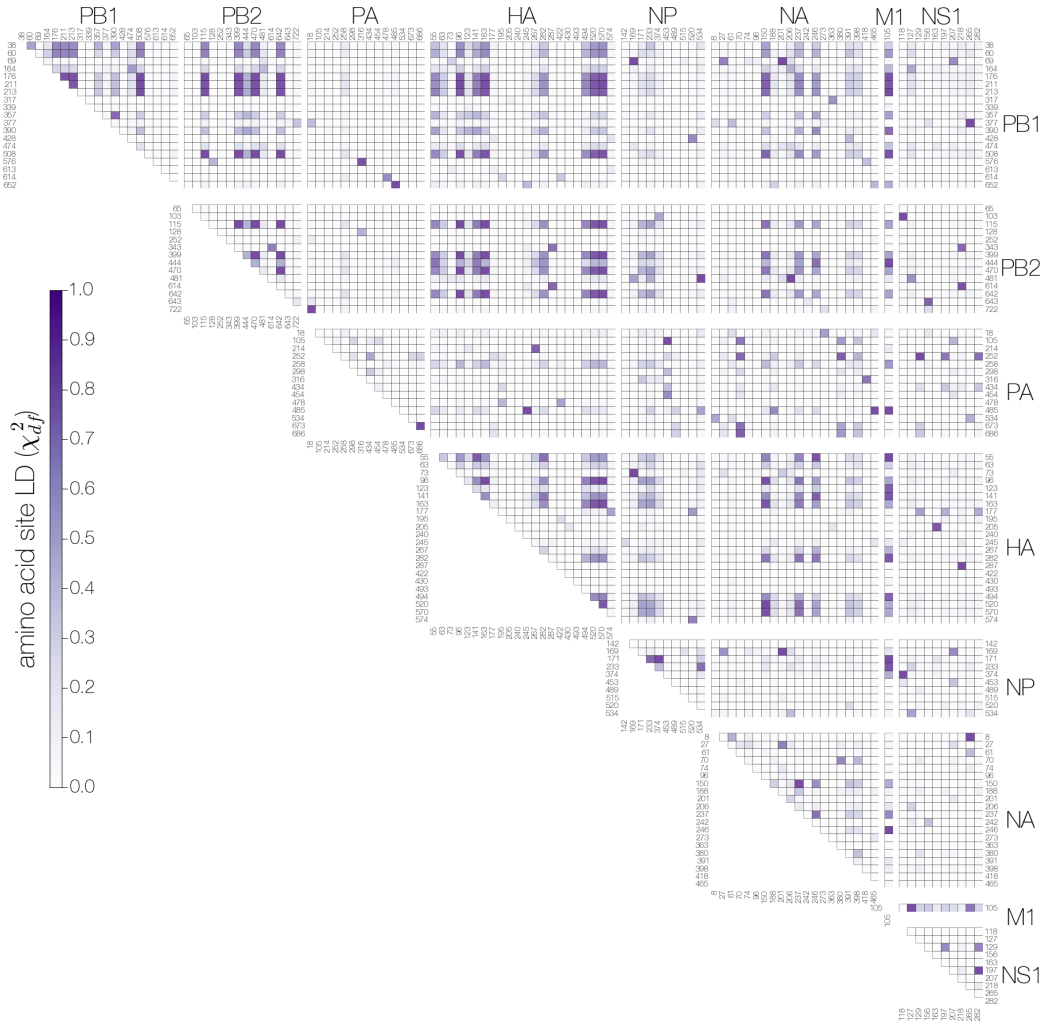


FIGURE B.3: **Heatmap of genome-wide linkage disequilibrium (χ^2_{df}) between polymorphic amino acid sites.** Patterns of LD across the genome suggest a network of reciprocally linked amino acid sites on PB1, PB2, HA and, to some extent NA, proteins. Proximity of sites on heatmaps might not correspond to proximity of sites within proteins.

segments differing by a single synonymous difference is highly efficient (Marshall et al., 2013). We note, however, that because of phylogenetic uncertainty our estimate of SPR distance might simply lack power. Comparisons between independent analyses of the same segments yield distances that are comparable to distances between different segments (Figures B.6 and B.7), suggesting that phylogenetic uncertainty is making a considerable contribution to our estimates of approximate SPR distances. Still, we find that independent replicates from the same segment (Figure B.7) show lower SPR distances that comparisons between segments (Figure B.6), suggesting that phylogenetic noise is not completely overwhelming reassortment signal. In addition, SPR distances

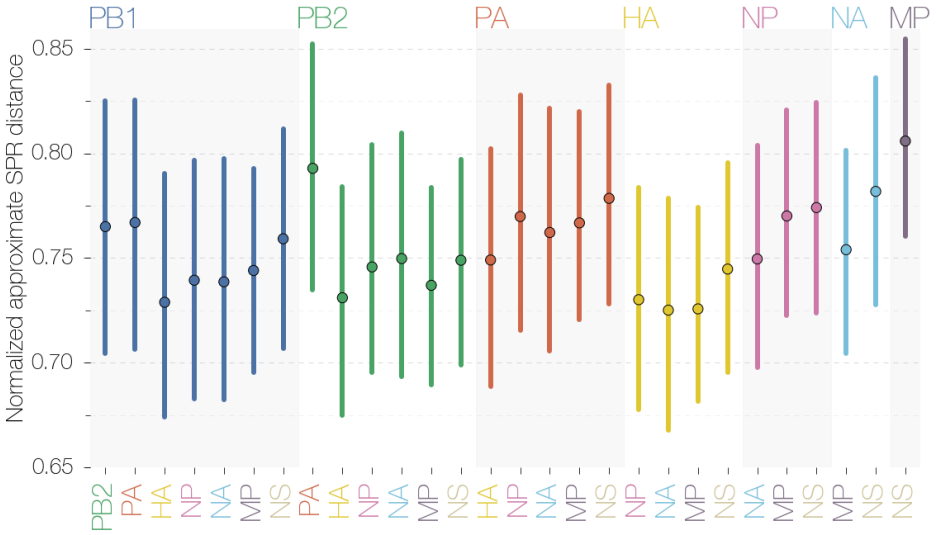


FIGURE B.4: **Normalized approximate SPR distances between pairs of segments.** Following the normalization procedure approximate SPR distances are similar across all pairwise comparisons. We interpret this as lack of evidence for small-scale topological similarities between trees of all segments, which we expect to arise if any two segments were being co-packaged and co-reassorted. All vertical lines indicating uncertainty are 95% highest posterior densities (HPDs).

themselves can only approximate (and underestimate) the actual numbers of reassortments. Thus we caution against over-interpreting Figure B.4. Although there might be concern about using approximate, rather than exact, SPR distances we do estimate exact SPR distances for a limited number of segment pairs - PB1, PB2 and HA - and find that after normalization exact and approximate SPR distances are not significantly different (Figures B.11–B.13).

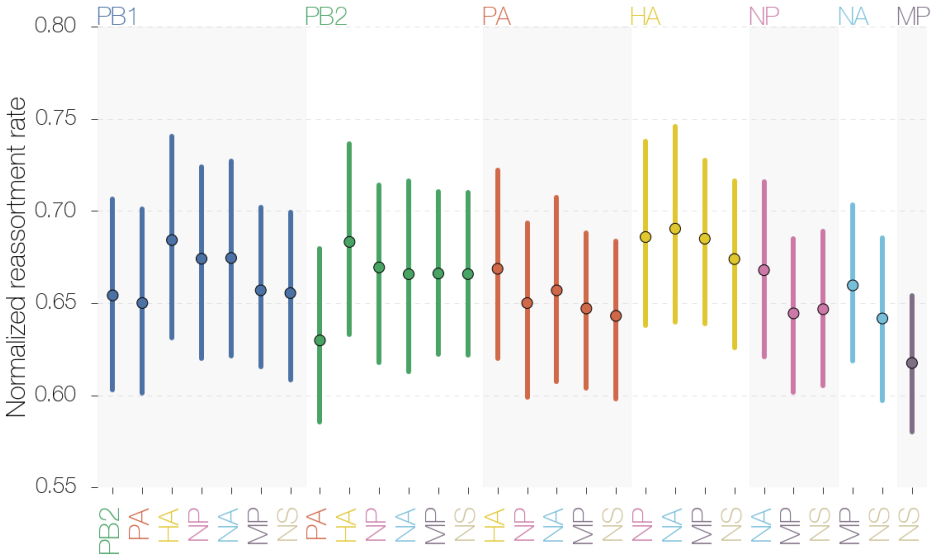


FIGURE B.5: **Normalized reassortment rate** Reassortment rate is calculated as approximate number of SPR moves per sum of total time in both trees.

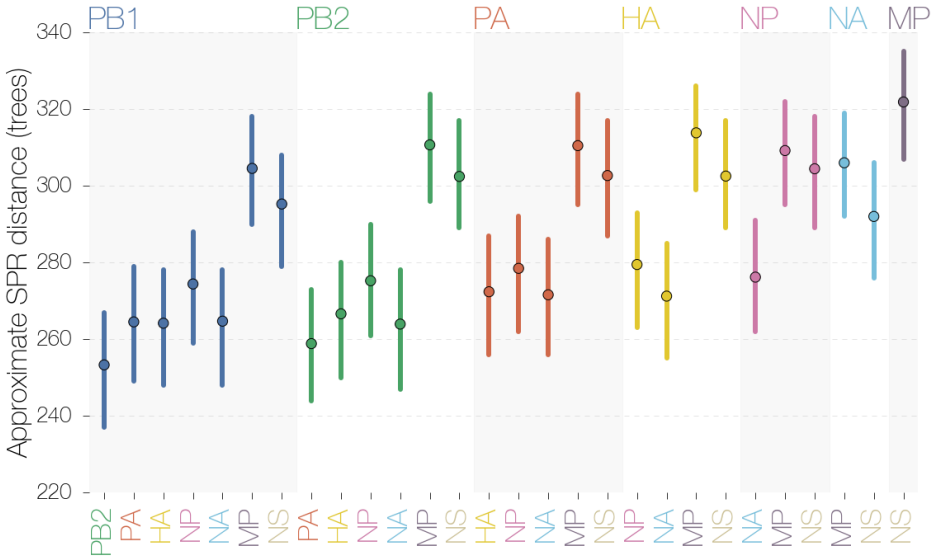


FIGURE B.6: **Approximate SPR distances between all pairs of trees of segments.** There is a visible trend where comparisons between shorter segments have larger SPR distances, consistent with decreasing tree topology stability over the course of MCMC for shorter segments.

FIGURE B.7: **Approximate SPR distances between replicate trees of each segment.** Approximate SPR distances between replicates of MP and NS trees are much higher (≈260) than any other segments, suggesting greater variability in tree topology over the course of MCMC. SPR distances between replicates of most other segments are ≈200.

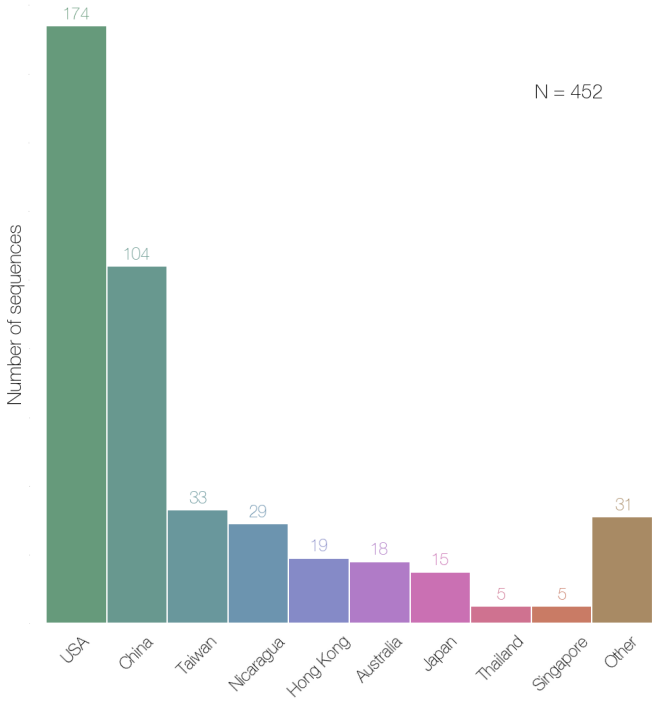


FIGURE B.8: **Geographic distribution of sequences in the primary dataset.** Sequences were assigned to the "other" category if there were less than 5 sequences from that country. Most of the genomes in the primary dataset were sampled in the USA.

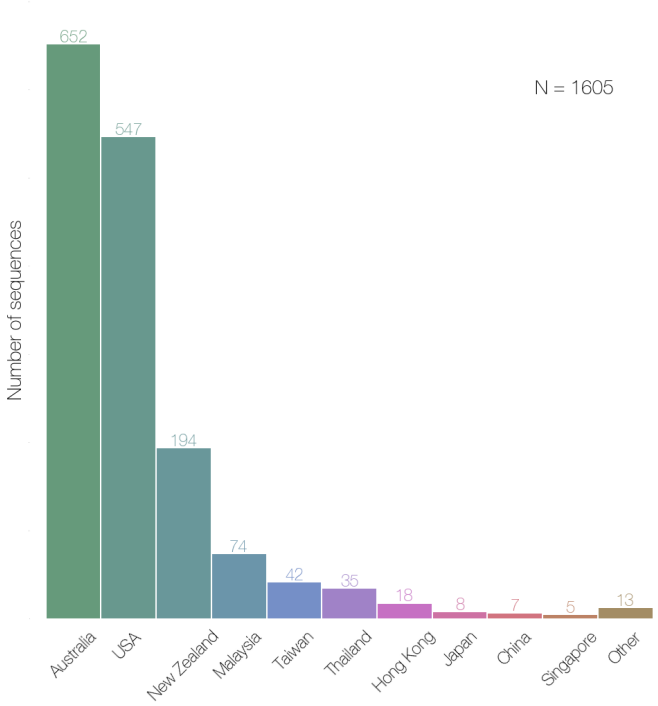


FIGURE B.9: **Geographic distribution of sequences in the secondary dataset.** Sequences were assigned to the "other" category if there were less than 5 sequences from that country. Most of the genomes in the secondary dataset were sampled in Australia.

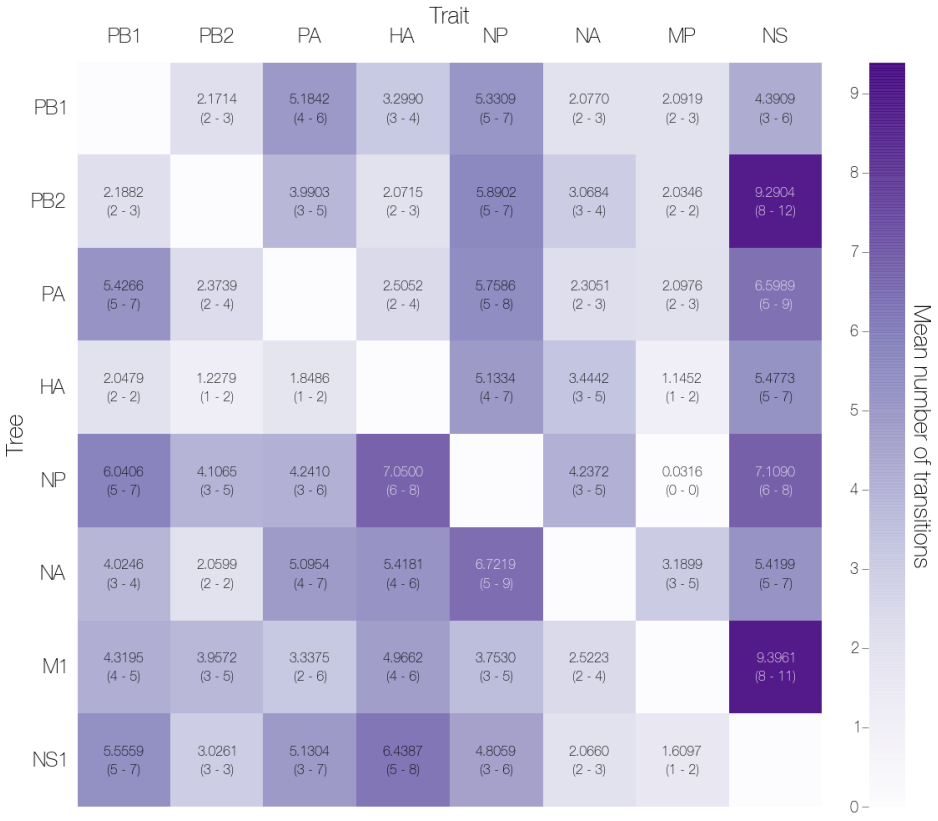


FIGURE B.10: **Numbers of trait transitions in each tree.** The numbers shown are the mean inferred number of trait transitions (minus one to account for the initial Vic–Yam split) in a given tree and trait combination. The numbers in brackets correspond to 95% highest posterior density intervals. Transitions may not be independent when more than one segment reassorts at the same time.

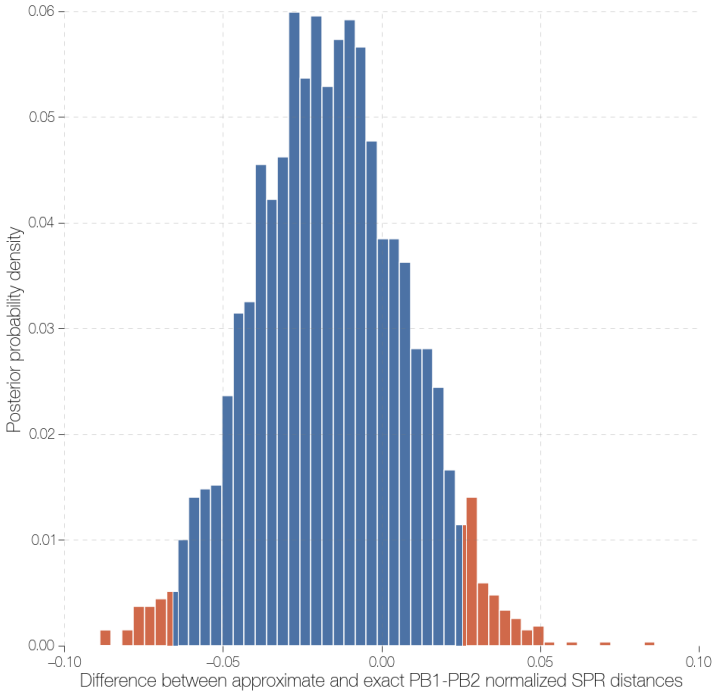


FIGURE B.11: **Distribution of differences between exact and approximate PB1-PB2 SPR distances after normalization.** 95% HPD interval (blue) overlaps zero, suggesting no evidence of differences between approximate and exact SPR distances following normalization.

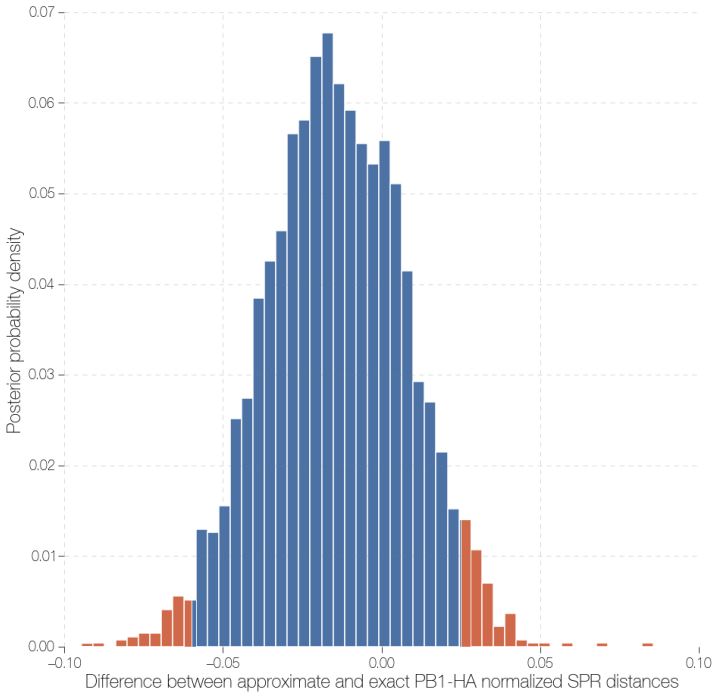


FIGURE B.12: **Distribution of differences between exact and approximate PB1-HA SPR distances after normalization.** 95% HPD interval (blue) overlaps zero, suggesting no evidence of differences between approximate and exact SPR distances following normalization.

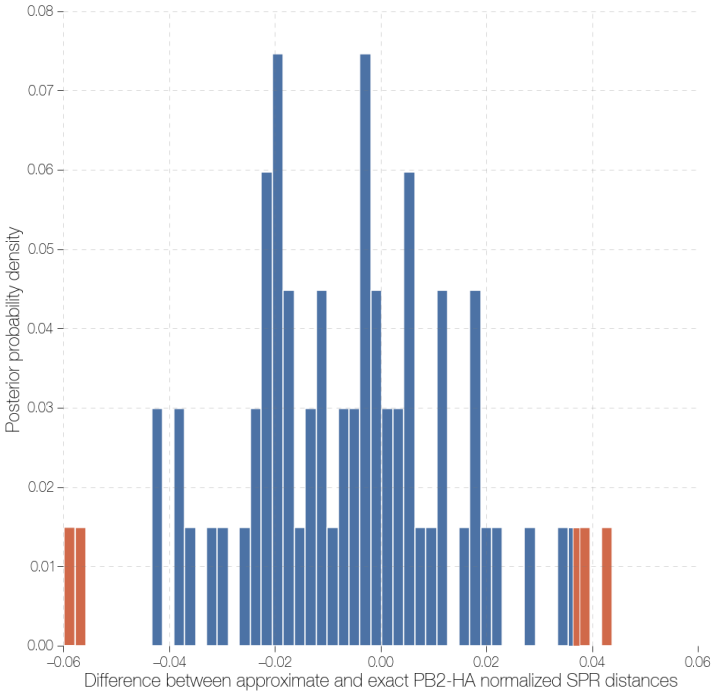


FIGURE B.13: **Distribution of differences between exact and approximate PB2-HA SPR distances after normalization.** 95% HPD interval (blue) overlaps zero, suggesting no evidence of differences between approximate and exact SPR distances following normalization. Due to excessively long computation time of exact SPR distances between PB2 and HA trees few comparisons were made.

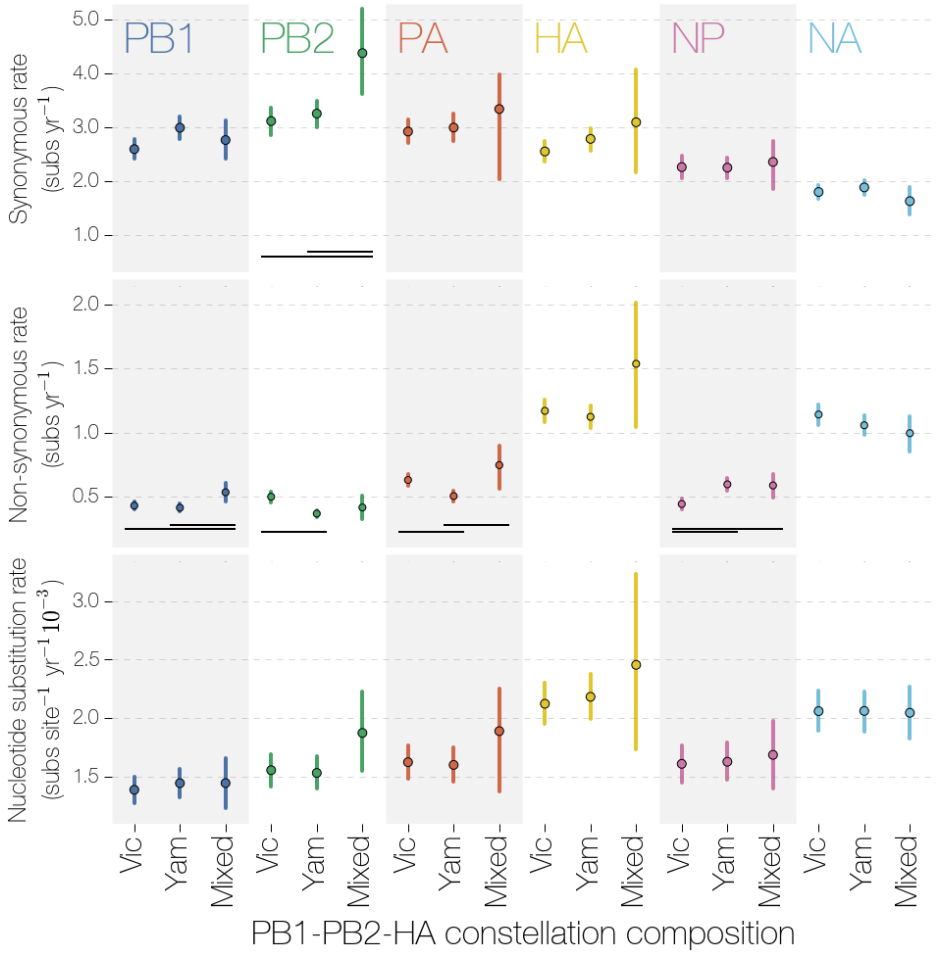


FIGURE B.14: **Synonymous, non-synonymous and nucleotide substitution rates in segments under different PB1-PB2-HA complexes.** Evolutionary rate dissimilarities under Vic and Yam PB1-PB2-HA complexes are not systematic and appear negligible. Synonymous and non-synonymous rates were calculated by dividing the sum of all substitutions of a given class by the total amount of evolutionary time under each PB1-PB2-HA constellation. Nucleotide rates were calculated by multiplying the inferred nucleotide substitution rate on each branch by the branch length, then dividing this by the total amount of evolutionary time under each PB1-PB2-HA constellation. Vertical bars indicating uncertainty are 95% HPDs, black bars indicate 95% HPDs that do not overlap.

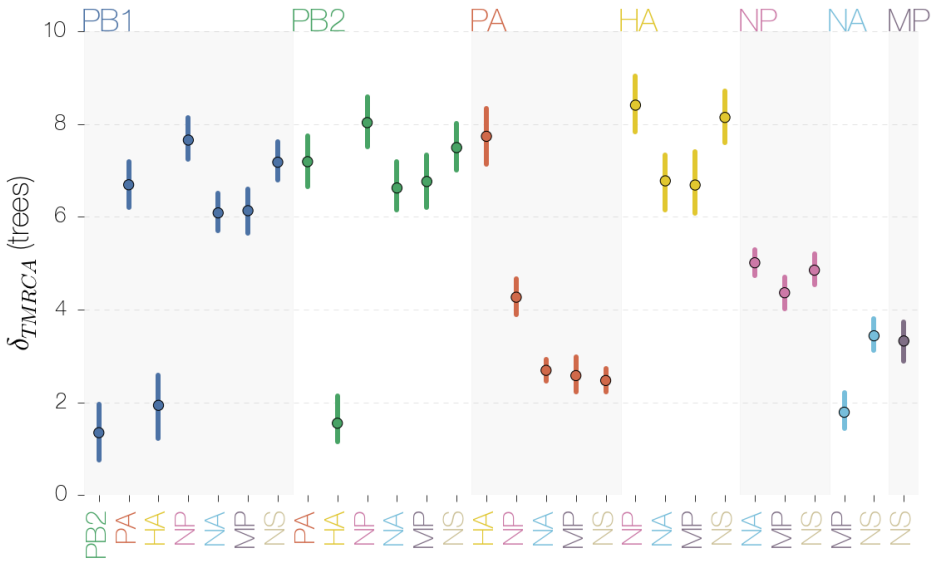


FIGURE B.15: δ_{TMRCA} **between all pairs of trees of segments.** δ_{TMRCA} between trees of segments reveal that tip pairs in PB1, PB2 and HA trees have very similar TMRCAs. The upper tail of the 95% HPD (HPDs are represented as vertical lines) interval of δ_{TMRCA} values for PB1-PB2-HA and MP-NA trees do not exceed 3 years.

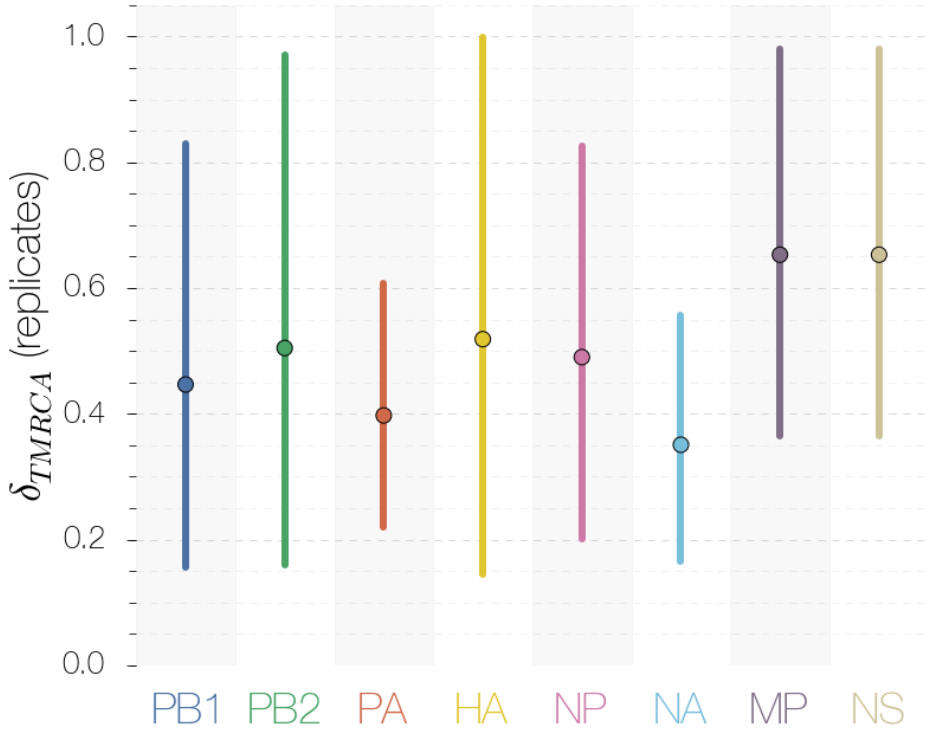


FIGURE B.16: δ_{TMRCA} **between replicate trees of each segment.** δ_{TMRCA} values between independent analyses of each segment show that mean δ_{TMRCA} values rarely exceed 1 year.

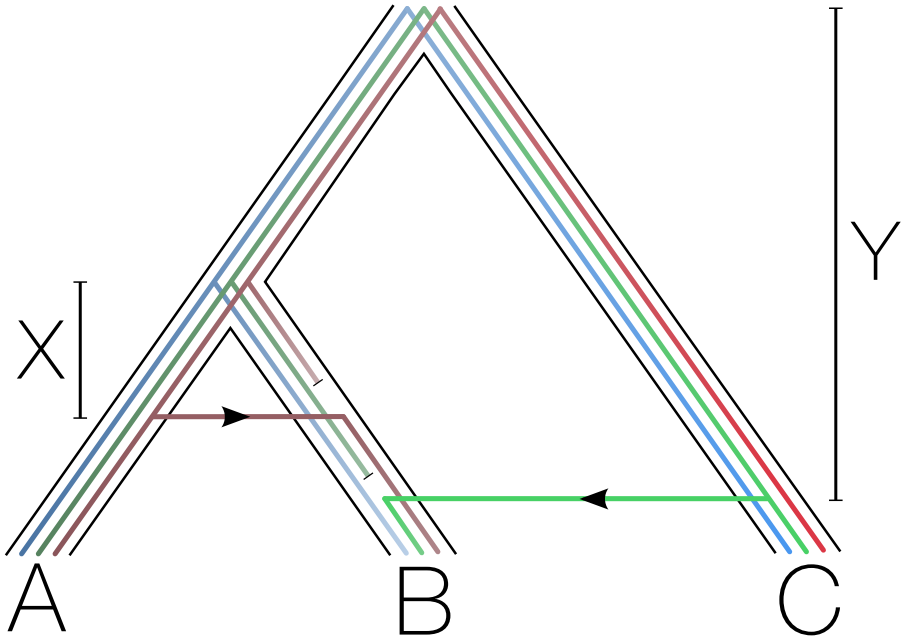


FIGURE B.17: **Calculating δ_{TMRCA} from a species tree perspective.** Consider an organism that has diverged into 3 taxa (A, B, C) with a genome comprised of 3 segments (blue, green and red). Due to reassortments taxa A and B share a slightly more recent TMRCA in the red segment, likewise for taxa B and C in the green segment. By comparing differences in TMRCAs between taxa A-B, A-C and B-C in blue, red and green segments we would find that the red segment has a lower 'reassortment distance' (X) than the green segment (Y). In the absence of reassortment we expect every segment in the genome to have the same tree, *i.e.* the tree of every segment should recapitulate the 'virus' tree (analogous to 'species' trees in diploid population genetics), including the dates of nodes. Due to population bottlenecks influenza viruses go through each year we expect strains isolated at any given time to have descended from a single recent virus genome. This descent from a single genome should therefore be reflected in the TMRCAs of all segments, the only exception being reassortment, which will dramatically alter the TMRCAs of the reassorted segment tree with respect to the background onto which it is reassorting.

APPENDIX C

REASSORTMENT BETWEEN CLOSELY RELATED INFLUENZA B LINEAGES AND THEIR DISPERSAL IN SCOTLAND

Methods supplement

Virus sequencing

This study used respiratory samples collected at hospital and primary care settings from throughout Scotland and referred to the Specialist Virology Centre at the Glasgow Royal Infirmary (GRI), and in South East Scotland referred to the Specialist Virology Centre at the Royal Infirmary of Edinburgh (RIE). Handling and sequencing of specimens for the study was carried out as part of Health Protection Scotland (HPS) funded Influenza reference work or in accordance with local ethical approval (South East Scotland Scottish Academic Health Sciences Collaboration (SAHSC) Human Annotated BioResource reference No.10/S1402/33). Nucleotide sequences of the influenza B virus HA and NA segments were obtained from randomly chosen archived nucleic acid samples. cDNA was generated from 5μl RNA using random hexamers as primers in 20μl reactions containing 15 U AMV reverse transcriptatse (Promega) with incubation at 37°C for 60 minutes. Hemi-nested PCR was carried out on 2μl of this cDNA in 20μl reactions using GoTaq DNA polymerase (Promega) with the primers 65:
(5'-GTACTACTCATGGTAGTAACATCC-3')
and 1364:
(5'-TTTCGTTRTGSAGTTCATCCATSGC)
for the first round and primers 65 and 1189:
(5'AATCATKCCTTCCCAKCCKCCT-3')
for the second round each for 30 cycles with the parameters 94°C for 18 sec, 50°C for 21 sec and 72°C for 90 sec. Primers for NA were pairs 8:
(5'-GCAGAGCATMTTCTCAAAACTGARG-3')
and 935:
(5'-GTTRTCTCTRCARGCACATTCTATGG-3'),
and 667:
(5'-ATATGGAGAAGCATAYACTGACACATA-3'
and 1496:
(5'-AAYAGGAACAAAGGGTTTAGAACAGA-3') for the first round
and pairs 47:
(5'-ATGAACAATGCTACCYTCAACTRTAC-3')

and 829:

(5'-CCCTCTCGAATCTTAAGRAATCTG-3'),

and 715:

(5'-TCCTAAGRACACAAGAAAGTGCCTG-3')

and 1439:

(5'-AACRCCTGTGACAGTGTCCCA-3'). PCR products were sequenced on both strands using BigDye reactions separated on an ABI 3730 machine. Nucleotide sequences were assembled, and aligned with reference sequences using SSE v1.1 (Simmonds, 2012).

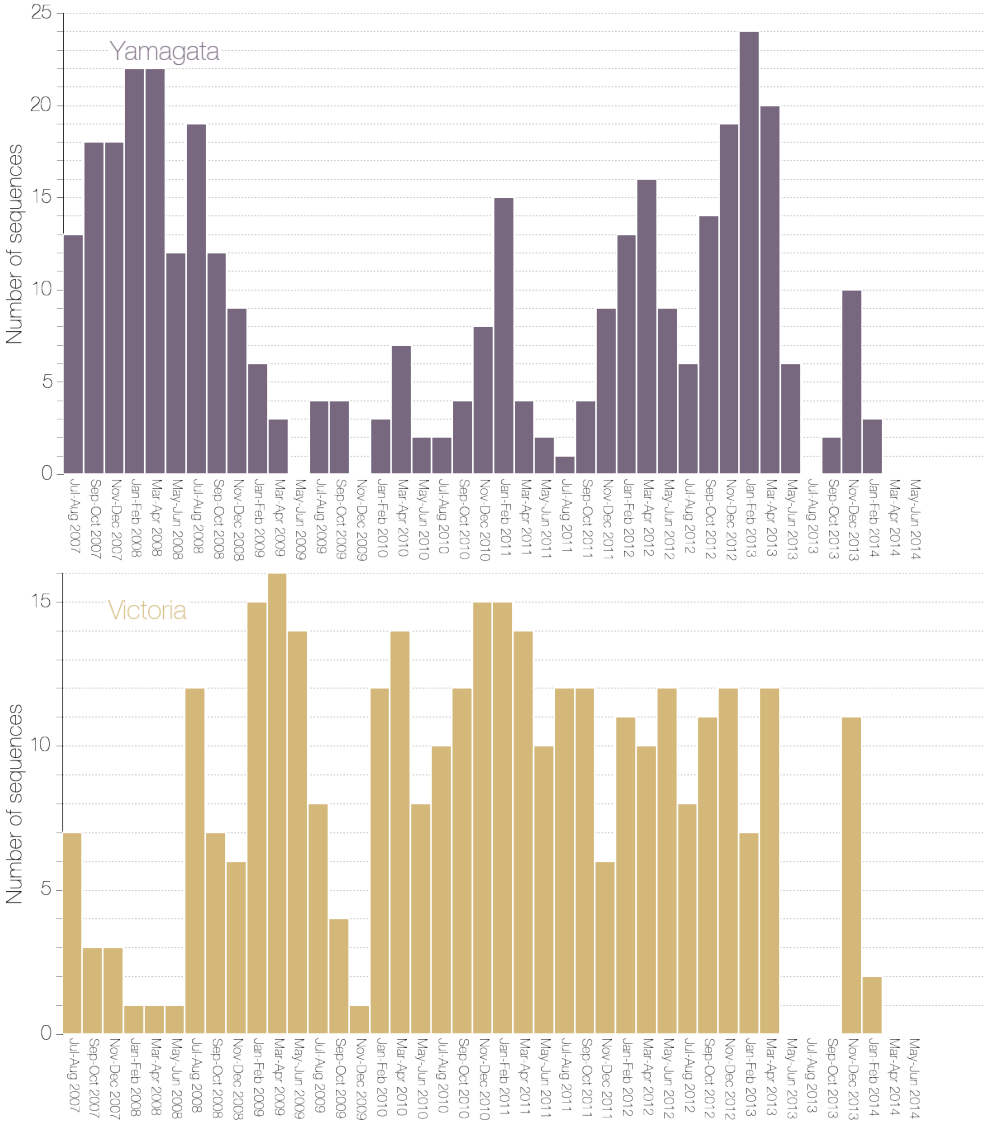


FIGURE C.1: **Temporal sampling within dataset 1.**

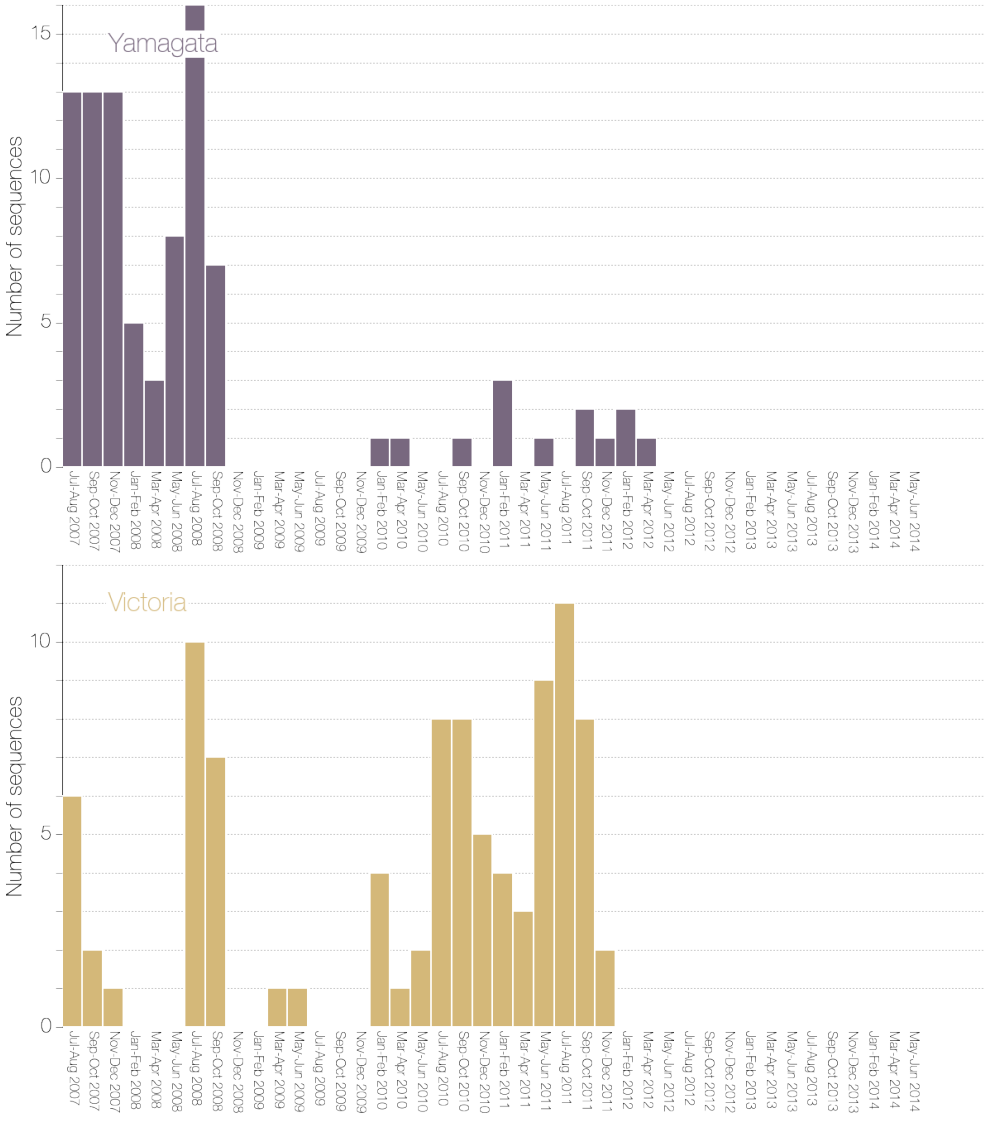


FIGURE C.2: **Temporal sampling within dataset 2.**

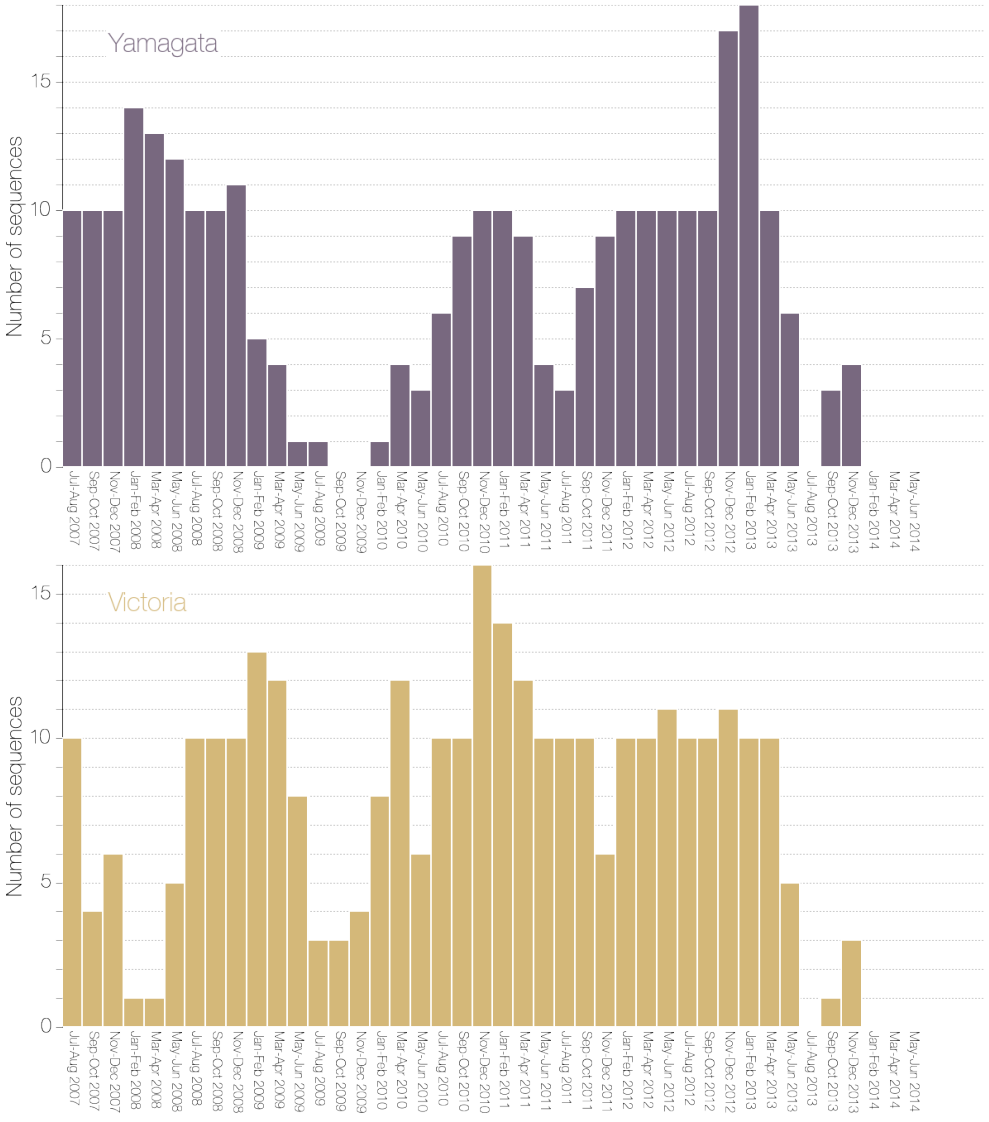


FIGURE C.3: **Temporal sampling within dataset 3.**

APPENDIX D

ORIGINS OF INFLUENZA D VIRUS

nucleotide model (segment-wise)	demographic model	log(BF)
joint	constant	66.588
joint	exponential	65.850
joint	**logistic**	**67.528**
independent	constant	0.682
independent	exponential	0.000
independent	logistic	1.240

TABLE D.1: **Marginal likelihood estimates for different models tested.** Marginal likelihoods estimated by path sampling are expressed as log Bayes factor in comparison to the worst-performing model: independent nucleotide substitution models for each segment and an exponential demographic model. The best fitting model is in bold, where a nucleotide substitution model is shared by all segments and logistic growth is the preferred tree prior.